Development of a "nature run" for observing system simulation experiments (OSSEs) for snow mission development

Melissa L. Wrzesien[1,2], Sujay Kumar[1], Carrie Vuyovich[1], Ethan D. Gutmann[3], Rhae Sung Kim[1,4], Barton A. Forman[5], Michael Durand[6], Mark S. Raleigh[7], Ryan Webb[8,9], Paul Houser[10]

[1]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD
[2]ESSIC, University of Maryland, College Park, MD
[3]National Center for Atmospheric Research, Boulder, CO
[4]GESTAR, Universities Space Research Association, Columbia, MD
[5]Department of Civil and Environmental Engineering, University of Maryland, College Park, MD
[6]School of Earth Sciences and Byrd Polar and Climate Research Center, Ohio State University, Columbus, OH
[7]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR
[8]Department of Civil, Construction, and Environmental Engineering, University of New Mexico
[9]Center for Water and the Environment, University of New Mexico, Albuquerque, NM 87131 USA
[10]Department of Geography and Geoinformation Sciences, George Mason University, Fairfax, VA

Corresponding author: Melissa Wrzesien, melissa.l.wrzesien@nasa.gov

1. Abstract

Snow is a fundamental component of global and regional water budgets, particularly in mountainous areas and regions downstream that rely on snowmelt for water resources. Land surface models (LSMs) are commonly used to develop spatially distributed estimates of snow water equivalent (SWE) and runoff. However, LSMs are limited by uncertainties in model physics and parameters, among other factors. In this study, we describe the use of model calibration tools to improve snow simulations within the Noah-MP LSM as the first step in an Observing System Simulation Experiment (OSSE). Noah-MP is calibrated against the University of Arizona (UA) SWE product over a Western Colorado domain. With spatially varying calibrated parameters, we run calibrated and default Noah-MP simulations for water years 2010-2020. By evaluating both simulations against the UA dataset, we show that calibration decreases domain averaged temporal RMSE and bias for snow depth from 0.15 to 0.13 m and from -0.036 to -0.0023 m, respectively, and improves the timing of snow ablation. Increased snow simulation performance also improves estimates of model-simulated runoff in four of six study basins, though only one has statistically significant improvement. Spatially distributed Noah-MP snow parameters perform better than default uniform values. We demonstrate that calibrating variables related to snow albedo calculations and rain-snow partitioning, among other processes, is a necessary step for creating a nature run that reasonably approximates true snow conditions for the OSSEs. Additionally, the inclusion of a snowfall scaling term can address biases in precipitation from meteorological forcing datasets, further improving the utility of LSMs for generating reliable spatiotemporal estimates of snow.

2. Introduction

Snow is a critical part of global and local water budgets, particularly in watersheds with headwaters in mountainous regions (Viviroli et al., 2007; Immerzeel et al., 2020). Millions of people around the world rely on snowmelt-derived runoff (Barnett et al., 2005; Li et al., 2017), especially in semi-arid regions. Despite being an integral component of global and regional water balances, estimating mountain snow accumulation remains one of the largest challenges of snow hydrology (Bormann et al., 2018; Dozier et al., 2016). While some mountain ranges have relatively dense in situ networks, other areas lack observations (Dozier et al., 2016), limiting techniques for interpreting point observations across a larger scale. Beyond in situ observations, remote sensing offers the ability to observe snow extent from space (Hall et al., 2002), but estimating snow water equivalent (SWE) to understand the water content of the snowpack remains a significant challenge, particularly in the mountains (Lettenmaier et al., 2015; Nolin, 2010; Takala et al., 2011; Vuyovich et al., 2014).

Due to limited in situ networks and uncertainty in remotely sensed observations, models are a practical alternative for developing spatiotemporal estimates of snow depth and SWE across large regions. Model intercomparison efforts have helped to identify important processes to improve simulating snow (Essery et al., 2009; Etchevers et al., 2004; van den Hurk et al., 2016; Krinner et al., 2018; Rutter et al., 2009), such as multi-layer snowpack. While snow models often have complex physics and parameterizations, resulting in accurate simulations of snow compared to in situ observations (Dutra et al., 2012; Etchevers et al., 2004), such processes are often too computationally complex for land surface models (LSMs) designed to run over large geographical areas. Additionally, snow models are typically focused only on modeling the snowpack processes whereas LSMs also enable the linkages to the water, energy, and carbon cycle processes. Though LSMs allow for simulations across a range of spatial and temporal scales in a computationally efficient manner, the relatively simple nature of their conceptual formulations and model parameterizations, as compared to complex process models, increases the uncertainties of their predictions. Further, biases in model forcing data, particularly precipitation, are a major driver of model error (Raleigh et al., 2015; Schmucki et al., 2014; Henn et al., 2018), and studies suggest that reanalyses, which are often used for model meteorological forcing, underestimate precipitation in mountainous areas (Henn et al., 2018; Enzminger et al., 2019; He et al., 2019). Such limitations are well documented in the literature, where it has been suggested that common LSMs, such as the Noah LSM with multiple parameterization options (Noah-MP; Niu et al., 2011), underestimate snow mass (Chen, Liu et al., 2014; Kumar et al., 2019; Xia et al., 2017; Chen, Barlage, et al., 2014). Despite these issues, LSMs are an essential tool for producing multi-year estimates of snow accumulation over continental or global study domains.

To reduce biases, models are often calibrated against reliable observation-based datasets (e.g. Ahl et al., 2008; Franz & Karsten, 2013; Henn et al., 2016; Rutter et al., 2009). Calibration has a long history in operational snow modeling (e.g. Turcotte et al., 2017; Franz et al., 2008) and previous intercomparison projects explicitly considered the performance of calibrated vs. non-calibrated models (Rutter et al., 2009; Essery et al., 2009). Often in snow and hydrological modeling, simulations are calibrated against discharge for improving model performance (Franz and Karsten, 2013; Hay et al., 2006; Ahl et al., 2008; Turcotte et al., 2017;). More recently, efforts have aimed to improve snow estimation by calibrating against SWE (Chen et al., 2017; Franz et al., 2010), snow-covered area (Franz and Karsten, 2013; Parajka and Bloschl, 2008), or multi-objective strategies that include two or more calibration variables (Nemri and Kinnard, 2020; Parajka et al., 2007; Chen et al., 2017; Franz and Karsten, 2013).

The performance of a calibrated model, however, will depend upon parameter selection for use during calibration, and complex LSMs such as Noah-MP have hundreds of parameters throughout the model code, some that are hard-coded to spatially uniform values. Cuntz et al. (2016) examined over 100 Noah-MP parameters, dozens of which are hard-coded into the LSM, and showed that simulated surface runoff is sensitive to almost all selected snow parameters; the authors conclude that it is necessary to expose some of the hard-coded parameters during calibration in order to improve model performance. Similarly, Mendoza et al. (2015) discussed that hard-coding parameters diminishes model agility; they identify several important hard-coded snow parameters that are treated as spatially uniform constants but in actuality likely vary through both time and space.

Here we calibrate Noah-MP against SWE estimates from the University of Arizona gridded observation-based snow data product (here referred to as UA; Zeng et al., 2018) in an effort to address dry biases in Noah-MP and improve snow estimation. We evaluate the impact of calibration on simulation of snow mass in a mountainous region. Since calibration will have implications beyond snow-related variables, we also examine impacts to other hydrologic processes, including runoff. The overarching motivation for the calibration is to produce a Noah-MP simulation that better approximates snow conditions through improvements to snow depth and SWE.

We aim for the calibrated simulation to be used as the “nature run” (NR) in a forthcoming snow-focused Observing System Simulation Experiment (OSSE). OSSEs are data assimilation experiments, performed to evaluate the type and impact of data to be collected from proposed missions and to enable the assessment of the utility from competing mission designs and design configurations (Garnaud et al., 2019; Crow et al., 2001, 2005; Wang et al., 2008; Nearing et al., 2012). Further, these experiments help to quantify the utility of observations beyond the immediate variable of interest (e.g., the impact of assimilating snow information on other aspects of the water budget, such as streamflow). OSSEs are useful in developing assessments of proposed observational methods and can be performed in addition to field work, such as the extensive NASA SnowEx campaigns, for evaluating proposed sensors.

A NR is the foundational step of an OSSE, upon which the data assimilation experiments are built (see Figure S1 for general steps to an OSSE). Within an OSSE, the NR simulation is considered the “true” state of the variable of interest. Therefore, NRs are developed using a high-quality model and meteorological inputs and should not have large uncertainty. Synthetic observations are then generated from the NR, after accounting for sources of errors and uncertainty associated with the anticipated sensor. The synthetic observations are assimilated into an open loop model simulation, and the assimilated result is compared back to the original NR to understand how well the proposed sensor captures the “true” conditions. The quality of the NR, therefore, significantly impacts the conclusions made from the OSSE. Since previous studies highlight biases in LSMs related to snow depth and SWE estimation, it is critical to reduce LSM bias and uncertainty to assess how proposed technologies perform in a variety of environments. If the NR and resulting synthetic observations are biased low, for example, it will be difficult to understand how a proposed sensor observes deep snowpacks. While a NR is not expected to be a perfect simulation, if it has a known systematic negative bias for SWE and snow depth, the assimilation experiments may not provide much information for how a sensor performs in regions where models have larger uncertainty, such as deep snow and forested regions (Kim et al., 2021). The calibration procedure described below is the first and an essential step in an OSSE designed to test potential configurations for a snow mission.

In addition to producing an improved NR for the OSSE, we aim to address three research questions: 1.) Can calibration address known dry biases in LSMs that cause underestimation of snow accumulation? 2.) How does calibration impact streamflow, beyond the targeted snow variables? and 3.) Can calibration suggest areas of model configuration that need improvement, such as meteorological forcing data for use as model boundary conditions? We test whether Noah-MP with calibration pre-processing yields similar snow estimates as a higher resolution, computationally expensive and complex snow physics model (SnowModel). We introduce the study area and calibration procedure in Section 3 below. In Section 4, we report results from the calibration experiments, and in Section 5, we discuss implications and provide thoughts for future studies.

3. Data and Methods

3.1. Model Setup

We use the NASA Land Information System (LIS; Kumar et al., 2006; Peters-Lidard et al., 2007) for simulations over a western Colorado domain. The domain is selected to include sites from previous NASA SnowEx field campaign locations, including Grand Mesa and Senator Beck (Figure 1). LIS is a land surface modeling framework designed to be highly flexible, offering users choice of LSM, meteorological forcing, and assimilation of in situ and remotely sensed observations, among other options. Created to be computationally efficient, LIS can perform simulations over large regional and global domains. The central component of the LIS framework is the LSM selection; LIS offers several community-supported LSMs relevant to operations and research. Here we use Noah-MP version 4.0.1. Recent work demonstrates that Noah-MP has superior performance to the original Noah LSM for simulating snow (Chen, Barlage et al., 2014; Chen, Liu et al., 2014; Kim et al., 2021; Minder et al., 2016; Wrzesien et al., 2015) due to model physics updates, including a multilayer (three layer) snowpack. Table S1 lists the physics options selected for the Noah-MP simulation.

In the LIS framework, Noah-MP simulates both surface water and energy fluxes as they respond to meteorological boundary conditions supplied by LIS. Simulations are from September 2009 through July 2020 at 0.01° spatial resolution (~1 km) and use hourly meteorological forcing data from the North American Land Data Assimilation System phase 2 (NLDAS-2; Xia et al., 2012). LIS includes statistical downscaling procedures for matching meteorological data to the specified spatial resolution of the LSM. The 1/8° spatial resolution NLDAS-2 forcing data are downscaled to ~1 km through a bilinear spatial interpolation approach. The model was first spun up for 72 years beginning in January 1979 and running through January 2020 twice until the simulation begins in September 2009. We also simulate the same time period using the default parameters to understand how calibration impacts the Noah-MP results. We distinguish between the two simulations as Noah-MP-Cal and Noah-MP-Def to represent the calibrated and default configurations, respectively.

3.2. Noah-MP Parameter Calibration

Previous studies suggest that LSMs underestimate snow accumulation, particularly in mountains (Broxton, Zeng, et al., 2016; Wrzesien et al., 2017, 2018). A recent model intercomparison using an ensemble of LSM simulations from LIS highlighted the model disagreement and uncertainty of snow estimation over North America, including mountain areas (Kim et al., 2021). To improve Noah-MP simulations, we select 24 parameters for calibration (Table 1), based on previous sensitivity studies (Cuntz et al., 2016; Mendoza et al., 2015) and their relationship to modeled snow processes. In Noah-MP-Def, these parameters are either hard-coded, often to a single spatially uniform value, or provided in lookup tables that vary based on land or soil properties. In contrast, the results from calibration are spatially distributed parameters that can vary across the domain (Figure 2). In addition to 23 existing parameters within Noah-MP, we include a snowfall scale factor in the calibration. Precipitation underestimation will impact the snow simulation and lead to biases throughout the snow season. The inclusion of a snowfall scale factor allows us to target the uncertainty resulting from biases in precipitation forcing. All 24 parameters are explored in point scale and full domain tests, though only the parameters that are sensitive enough to warrant calibration are described in Section 4.1.

Noah-MP is calibrated against SWE estimates from the University of Arizona dataset (UA; Zeng et al., 2018) in an optmization approach. The UA data product provides SWE at 4 km spatial resolution over the conterminous United States (Zeng et al., 2018). Estimates are provided daily between 1981 and 2020. UA is based on the assimilation of in situ measurements of both SWE and snow depth (Broxton, Dawson, et al., 2016) and precipitation and temperature values from the PRISM dataset (Daly et al., 2000). UA has been evaluated against multiple datasets (Dawson et al., 2018), including airborne lidar measurements of snow depth. We note that any biases in UA SWE will likely be reflected in the calibrated parameters and the resulting simulations; however, such biases, especially in gridded observation-based data products like UA, are unavoidable.

We calibrate over water years 2007-2009. This period was selected by examining domain averaged SWE from water years 1982-2020 from the UA record. From comparisons of domain-wide average maximum SWE and depth, this period included average (2009), high (2008), and low (2007) snow conditions for the study region. Domain-wide average maximum SWE (snow depth) for water years 2007, 2008, and 2009 is 135.5 mm, 231.0 mm, and 162.4 mm, respectively, (524.0 mm, 850.1 mm, and 536.9 mm, respectively) vs. the long-term mean of 163.5 mm (618.5 mm).

For calibration, we use a genetic algorithm (GA), which is part of the LIS-Optimization and Uncertainty subsystem (Kumar et al., 2012). The GA is a common stochastic tool used in hydrology model optimization (Duethmann et al., 2014; Isenstein et al., 2015; Shafii & De Smedt, 2009; Wang, 1991; Yapo et al., 1998) and is designed to mimic biological evolution where the fittest of the population (i.e., parameter sets), as determined through comparison to an observational dataset, survive and move to the next generation. Within each generation, crossover and mutation operators are used to produce new parameter estimates and to introduce diversity in the parameter set. To ensure good solutions are not lost between generations due to either crossover or mutation operators, an elitism strategy is used, where the best solution is carried over to the next generation. Over many generations, the average fitness, which reflects the quality of the solution, tends to increase due to the selection of individuals that compare favorably to observations.

GAs aim to prevent overfitting through an ensemble approach and by introducing poor performing solutions through mutation operators. Since they do not rely on gradient information, GAs can handle local optima and discontinuities in the search space, unlike gradient search. Since GAs require an ensemble that must be run over several generations, they are computationally expensive. Running 50 generations of the GA with 30 ensemble members for three water years over the study domain requires a total running time over 480 hours, or over 20 days of continuous simulation, with 532 processors.

Within LIS, the GA does not provide estimates of parameter uncertainty. For estimating parameter uncertainty, variants of Markov Chain Monte Carlo methods such as Differential Evolution Monte Carlo (te Braak et al., 2008) would be required; however, algorithms such as these have a high computational cost, with run times an order of magnitude higher than GA (Harrison et al., 2012), making their implementation over a domain size such as ours difficult. Since the primary objective of this study is to produce a better snow simulation, a thorough investigation into the parameter uncertainty is omitted. More detail on GAs within the LIS framework is discussed by Kumar et al. (2012).

The GA results in calibrated values for a set of parameters that allow for the best match with observations. The range in parameter values for calibration (see Table 1) are either taken from the literature or allowed to vary +/- 20% of the default value, following Cuntz et al. (2016). As an objective function, we consider the squared difference between the observation and the model:

$$J_i = (d_i^o - d_i^m)^2 \qquad (1)$$

where d_i^o is snow depth from the observations (UA) for grid cell i and d_i^m is snow depth from the model (Noah-MP) for grid cell i. We minimize J_i for each grid cell i independently in the calibration, resulting in parameters that vary spatially (Figure 2). In contrast, Noah-MP-Def has spatially uniform parameters. UA, produced at 4 km, is rescaled to match the Noah-MP resolution through bilinear interpolation during calibration.

3.3. Evaluation Datasets

In addition to comparing Noah-MP estimates to UA, we evaluate snow simulations against a suite of independent datasets using the Land surface Verification Toolkit (LVT; Kumar et al., 2012). First, we compare snow depth across the full domain to the Snow Data Assimilation System (SNODAS; Carroll et al., 2001), which is an operational dataset available over the contiguous United States at approximately 1 km spatial resolution. Both Noah-MP simulations are evaluated against UA and SNODAS for the full analysis period of water years 2010-2020. UA and SNODAS are both reprocessed in LVT to match the spatial resolution of Noah-MP.

We also compare to snow depth measurements from the Global Historical Climatology Network (GHCN; Menne et al., 2012); the western Colorado domain includes 79 GHCN stations with snow depth observations. Stations within the domain include a range of elevations (1467-3422 m) with an average station elevation of 2349 m. This compares to the full Noah-MP domain with elevations ranging from 1399-4185 m and an average elevation of 2639 m; approximately 9% of GHCN stations within the domain have elevations > 3000 m, compared to 26% of the full domain. While GHCN stations undersample higher elevations within the western Colorado domain, they provide an additional evaluation dataset for snow depth. GHCN data are available for water years 2010-2016.

We also compare Noah-MP to datasets collected from the 2017 NASA SnowEx field campaign in Colorado. First, we evaluate Noah-MP against snow pit observations of snow depth and SWE from SnowEx (Elder et al., 2018) at Grand Mesa and Senator Beck, which were collected between February 6-25, 2017. For a spatial comparison, we evaluate Noah-MP snow depth against Airborne Snow Observatory (ASO) lidar observations of snow depth, which are produced at 3 m spatial resolution (Painter, 2018). Here we use ASO flights over Grand Mesa from February 8 and February 16; though other flights are available for the 2017 field campaign, other days either included artefacts from the lidar collection or excluded portions of the mesa.

In addition to observations from SnowEx, Noah-MP is evaluated against a SnowModel simulation over Grand Mesa for the 2017 campaign, as described in Webb et al. (2020). SnowModel is a widely used snow model that simulates distributed snow properties in space and time and can be configured to simulate a single or multi-layer snowpack (Liston and Elder, 2006a; Liston and Sturm, 1998). SnowModel is designed to include four interconnected models: MicroMet for processing and downscaling meteorological forcing data (Liston and Elder, 2006b), EnBal for calculating the energy balance of the snowpack, SnowPack for simulating the snowpack in space and time, and SnowTran-3D for computing redistribution of snow due to wind (Liston and Sturm, 1998; Liston et al., 2007). Webb et al. (2020) configure SnowModel to simulate a single layer snowpack over Grand Mesa for the 2016-2017 water year to coincide with the SnowEx field campaign in February 2017. They use station observations as meteorological forcing data, including data from the Grand Mesa Study Plot (Skiles, 2018), four SnowEx campaign weather stations, and three nearby Snow Telemetry (SNOTEL) sites. SNOTEL sites provide temperature and precipitation observations, and all other stations provide temperature, wind speed/direction, humidity, and radiation. No adjustment of precipitation or other forcing data were made, and SnowModel simulations were independent of any snow observations. Elevation data were from the 1/3 arc-second USGS National Elevation Dataset, while vegetation data were taken from 30m USGS LANDFIRE v.1.4 Existing Vegetation Type data (Rollins, 2009) and reclassified to SnowModel vegetation types. Webb et al. (2020) ran SnowModel at multiple spatial resolutions, but here we consider SWE and snow depth outputs from their 30 m simulation. Webb et al. (2020) provide additional information on the SnowModel configuration and evaluation.

For spatial evaluations against both ASO and SnowModel, we calculate the Spatial Efficiency (SPAEF; Koch et al., 2018; Demirel et al., 2018), which combines histogram matching, spatial correlation coefficient, and spatial variability error to evaluate spatial patterns. SPAEF is defined as:

$$SPAEF = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \qquad (2)$$

where $\alpha = \rho(obs, mod)$, $\beta = \frac{\frac{\sigma_{mod}}{\mu_{mod}}}{\frac{\sigma_{obs}}{\mu_{obs}}}$, and $\gamma = \frac{\sum_{j=1}^{n} \min(K_j, L_j)}{\sum_{j=1}^{n} K_j}$. Here α is the Pearson correlation coefficient between the observation (ASO lidar or SnowModel simulation) and the model (Noah-MP), β is the fraction of the coefficient of variation, which represents spatial variability, and γ is the histogram intersection for the histogram of the observation, K, and the histogram from the model, L (Swain and Ballard, 1991). SPAEF has an optimal value of 1.

For streamflow, we compare to natural flow estimates for four basins in the Upper Colorado River Basin (UCRB) that lie completely within the model domain (see Table 6). Natural flow estimates are from the Bureau of Reclamation and are available monthly between 1901 and 2018 (Prairie and Callejo, 2005). We also compare to daily, unregulated streamflow for two basins from the Catchment Attributes and Meteorology for Large Sample Studies (CAMELS; Newman et al., 2015; Newman et al., 2014) dataset. We only use streamflow observations between 2009 and 2014 for the two CAMELS basins, and the daily streamflow has been processed into monthly averages. Since Noah-MP does not include human management on streamflow networks, we cannot compare model-simulated runoff to streamgage observations, due to water diversions, dams, and other water management practices. Instead, we compare monthly grid-cell generated runoff – the summation of surface runoff and subsurface runoff – to monthly observations over small unmanaged basins and to estimated natural flow (i.e., runoff in the absence of human management) in larger basins. Using total runoff at monthly scales as a proxy to streamflow is a valid assumption (Chow, 1964) and a strategy used in other studies (e.g., Koster et al., 2010). We evaluate monthly streamflow with Nash-Sutcliffe Efficiency metrics (NSE; Nash & Sutcliffe, 1970), where a perfect fit with observations has NSE = 1, and NSE > 0 indicates the model has better predictive skill than the mean of the observations.

4. Results

4.1. Calibration

We initially run point-scale calibration tests with 23 selected parameters from the snow modules within Noah-MP (Table 1). Noah-MP-Cal generally improved the snow ablation timing in spring months relative to Noah-MP-Def. However, maximum snow conditions remained largely underestimated, particularly for sites with deep snowpack (not shown). After implementation of a snowfall scaling factor, described below in equation 5, as an additional calibration parameter, test simulations resulted in snow depths in better agreement with UA estimates. Therefore, for calibration over the full domain, we include 24 spatially variable parameters: 23 from Noah-MP and an additional snowfall scale term (Figure 2).

Though we include 24 parameters in the GA procedure, only 11 were sensitive to calibration. We determine that 13 are not sensitive because they do not demonstrate any noticeable spatial patterns such as those reported in Figure 2 and instead calibrated values have noisy spatial patterns (see Supplemental Figure S2). Some of the 11 selected parameters have regions of noisy artificial patterns in regions of the domain that were insensitive to calibration, often in portions of the domain where less snow accumulates (Figure 2). Despite these regions, we look further into the 11 sensitive parameters. The first four parameters are used within the CLASS snow albedo scheme (Verseghy, 1991) and include minimum snow albedo ($MNSNALB$), maximum snow albedo ($MXSNALB$), the exponent in the snow albedo decay relationship ($SNDECAYEXP$), and the new snow mass required to cover old snow ($SWEMX$). These parameters are used in each time step to calculate snow albedo. First, the albedo of the snow cover for the new time step is determined as:

$$\alpha_s(t) = MNSNALB + [\alpha_s(t-1) - MNSNALB] * \exp\left[-\frac{SNDECAYEXP*\Delta t}{3600}\right] \quad (3)$$

where α_s is snow albedo at time step t or $t-1$ and Δt is the model time step. If new snow has fallen in an amount larger than $SWEMX$, snow albedo is refreshed to a value of $MXSNALB$.

The next group of calibration parameters relates to the rain-snow partitioning scheme used here, i.e., the Jordan (1991) scheme from the SNTHERM model. In this method, if air temperature is above the upper temperature limit ($T_U LIMIT$), all precipitation is rainfall. At air temperatures below the lower temperature limit ($T_L LIMIT$), all precipitation is snowfall. For temperatures between $T_L LIMIT$ and a middle threshold ($T_M LIMIT$), the fraction of precipitation that is frozen is a function of air temperature. At temperatures between $T_M LIMIT$ and $T_U LIMIT$, the fraction of precipitation that is frozen is set to 0.6. In the calibration procedure, $T_L LIMIT < T_M LIMIT < T_U LIMIT$.

The remaining four parameters are from different schemes throughout Noah-MP, and three were highlighted by Mendoza et al. (2015) as key parameters for model sensitivity. These include the exponent used in the snow depletion curve ($MFSNO$), liquid water holding capacity (SSI), and snow surface roughness length ($Z0SNO$). $MFSNO$ is used within Noah-MP to calculate the fractional portion of the grid cell that is snow covered, as shown in equation 4 below from Niu and Yang (2007):

$$f_{sno} = \tanh \frac{h_{sno}}{2.5 z_{0g} \left(\frac{\rho_{sno}}{\rho_{new}}\right)^{MFSNO}} \qquad (4)$$

where h_{sno} is snow depth, z_{0g} is the bare soil roughness length, ρ_{sno} is bulk density of snow, and ρ_{new} is the density of new snow, which is set to 100 kg/m^3. f_{sno} is used throughout Noah-MP to scale grid-cell calculations into snow-covered and non-snow-covered fractions, including within surface radiation calculations.

SSI and $Z0SNO$ are each used only once in the Noah-MP code. SSI is included in the calculation of snow layer liquid water, which determines the rate of exfiltration of snowmelt release from the bottom of the snowpack. $Z0SNO$ is used to calculate the surface roughness length for turbulent flux calculations over snow covered ground.

The final calibration parameter is the snowfall scaling term, $SNOWF\_SCALEF$, which was included to address uncertainty in precipitation forcing data. $SNOWF\_SCALEF$ is described as:

$$S = P * f_{ice} * SNOWF\_SCALEF \qquad (5)$$

where S is snowfall, P is total precipitation, and f_{ice} is the fraction of the precipitation that is frozen. The snowfall scale factor is applied to frozen precipitation to reduce the bias introduced from NLDAS-2. Other studies introduce a similar precipitation scaling factor in optimization or assimilation experiments. Smyth et al. (2020), who also used NLDAS-2 for model forcing data, use a snowfall correction factor to scale precipitation at their SNOTEL study sites across the western United States. In their work, the average snowfall correction factor is 1.64, indicating NLDAS-2 underestimates mountain snowfall by more than 50%. In Smyth et al. (2020) and here, NLDAS-2 snowfall is too low and must be scaled to larger values to produce realistic snow accumulation. Other studies have also included a correction factor to address biases in snowfall from meteorological data (Magnusson et al., 2017; He et al., 2011; Franz and Karsten, 2013). Errors in forcing data, particularly precipitation, have a large impact on snow modeling performance (Raleigh et al., 2015; Schmucki et al., 2014; Henn et al., 2018), and including a snowfall scaling term in the calibration procedure can help address this bias.

4.2. SWE and Snow Depth Evaluation

4.2.1. UA and SNODAS Comparisons

4.2.1.1. Full Domain Comparison

Figure 3 shows the time series of average SWE and average snow depth across the domain for Noah-MP-Cal, Noah-MP-Def, and the UA dataset. In nearly all cases, calibration results in more snow and later snowmelt. Occasionally, Noah-MP-Cal produces more snow accumulation than the UA dataset, such as in 2015 and 2017 (Figure 3). Over the 11-year simulation, Noah-MP-Cal has larger magnitudes of snow depth and SWE; average maximum SWE (depth) from Noah-MP-Cal is 166.7 mm (0.61 m), while average maximum SWE (depth) from Noah-MP-Def is 131.8 mm (0.52 m).

Spatially, Noah-MP-Cal produces greater April 1 SWE at higher elevations across the domain, averaged over the water year 2010-2020 simulation period (Figure 4). Estimates from Noah-MP-Def have similar domain-wide averages as Noah-MP-Cal (Figure 3), but the snow is less spatially variable. This is contrasted with Noah-MP-Cal where snow accumulation more closely follows local topography. We also compare Noah-MP-Cal and Noah-MP-Def to UA and SNODAS at six evaluation points throughout the domain that correspond to SnowEx field campaign sites (Table 2). At these points, Noah-MP-Cal generally has smaller biases and RMSE than Noah-MP-Def for the UA comparison (Table 3). Noah-MP-Cal also tends to perform better than Noah-MP-Def when evaluated against SNODAS (Table 3). Noah-MP-Cal has smaller bias and RMSE at all evaluation points except Fool Creek and Senator Beck, the two highest elevations stations. For a similar comparison but for SWE, see Table S2.

To compare Noah-MP-Cal and Noah-MP-Def against UA and SNODAS over the full domain, we first calculate the SPAEF (Equation 2) to evaluate spatial performance. Compared to UA, Noah-MP-Cal has a SPAEF of 0.799 and Noah-MP-Def has a SPAEF of 0.508. For SNODAS, Noah-MP-Cal also has a higher SPAEF metric: 0.722 vs 0.460 for Noah-MP-Def. For RMSE (Figure 5), higher elevations tend to have larger RMSE values, particularly for Noah-MP-Def compared to both SNODAS and UA. Noah-MP-Cal has high RMSE values in the central northern portion of the study domain. This area has much larger values of snow depth in Noah-MP-Cal than Noah-MP-Def, and the snowfall scale factor from calibration is high in the area (up to 2.5-3, compared to the domain average of 1.16), leading to increased precipitation and higher snow accumulations (discussed in Section 5). Aside from this anomalous region and an area in the southern portion of the domain, Noah-MP-Cal generally reduces the UA snow depth RMSE (Figure 5c), particularly at higher elevations. Averaged over the domain, Noah-MP-Cal has a slightly lower RMSE (0.13 m) than Noah-MP-Def (0.15 m) compared to UA (Table 3). Performance between Noah-MP-Cal and Noah-MP-Def is similar for SNODAS as for the UA comparison. Averaged over the full domain, Noah-MP-Def is in better agreement with SNODAS (RMSE of 0.18 m) than Noah-MP-Cal (RMSE of 0.19 m), though results are generally similar.

Similar to RMSE, we also compare temporal bias over the full domain (Figure 6). Noah-MP-Def has a negative bias for higher elevation grid cells compared to both UA and SNODAS. This suggests that Noah-MP-Def is underestimating snow accumulation in the mountains, highlighting the known dry bias of LSMs (e.g., Chen, Liu et al., 2014; Holtzman et al., 2020; Kumar et al., 2019; Wang et al., 2019; Xia et al., 2017). Noah-MP-Cal bias spatial patterns are similar between both UA and SNODAS, with a large positive bias in the central northern portion of the domain due to anomalously high values of snow depth. Averaged over the full domain, Noah-MP-Cal vs. UA has a bias of nearly zero (-0.0023 m), compared to Noah-MP-Def of -0.036 m (Table 3). For both UA and SNODAS comparisons, Noah-MP-Cal has more instances of positive bias at higher elevations (>3500 m), while these same grid cells in Noah-MP-Def tend to have negative biases. Noah-MP-Def underestimates snow accumulation at high elevations and calibration somewhat addresses these biases, though can result in too much snow in some regions.

4.2.1.2. Seasonal Comparison

During the accumulation season (December through February), calibration increases the domain averaged snow depth by almost 18%, from a -14.0% difference with Noah-MP-Def to a +1.4% difference with Noah-MP-Cal, relative to UA. RMSE also improves slightly from 0.162 m to 0.142 m. Similarly, for the peak snow season (March and April), calibration results in an improvement of snow depth percent difference from -24.8% (Noah-MP-Def) to -5.1% (Noah-MP-Cal). RMSE decreases from 0.269 m with Noah-MP-Def to 0.215 m with Noah-MP-Cal, a 20% improvement. Calibration results in large improvements for the ablation season (May through July), increasing the domain averaged snow depth by 45.4%. Noah-MP-Def mean snow depth is 31.6% less than the UA estimate, while Noah-MP-Cal is comparable to UA, only -0.5% smaller. RMSE decreases by over 12%, from 0.0981 m with Noah-MP-Def to 0.0863 m with Noah-MP-Cal. Across the full domain, calibration addresses the underestimation of snow throughout the full water year, though with slightly too much snow during the peak snow season.

At the grid cell scale, Noah-MP-Cal generally has more snow accumulation and a later end to the snow season than Noah-MP-Def, as shown in Figure 7 for Senator Beck.

Point scale evaluations have a better agreement between UA and Noah-MP-Cal, with RMSE declining by 4.23 cm for peak season. During the accumulation and ablation seasons, results are different, where Noah-MP-Def has smaller bias and RMSE. Noah-MP-Cal overestimates UA in the spring for several years (Figure 7a), with snow lingering longer than observed in UA for water years 2015, 2017, and 2019. Performance is similarly mixed at other study points (Table 4), where calibration may improve performance during all seasons (Cameron Pass, Niwot Ridge, Skyway/Grand Mesa) or may degrade performance, depending on the season (accumulation and peak for Fool Creek, ablation for Rock Creek, and accumulation and ablation for Senator Beck). Comparing SWE bias and RMSE over different seasons has similar results (Table S3).

4.2.1.3. Comparison over Vegetation Class

We next aggregate the 20 LIS land cover classifications into five broader groups – forest, shrubland, grassland, cropland, and barren (see inset in Figure 1 and see Table S4 for statistics) – and compare Noah-MP-Cal vs. Noah-MP-Def against both UA and SNODAS (Figure 8a,c). For average snow depth bias, Noah-MP-Cal performs better than Noah-MP-Def across land covers. Most comparisons have a negative bias, indicating that Noah-MP-Cal and Noah-MP-Def have less snow than either UA or SNODAS, though magnitude of the bias is generally smaller than 0.05 m. The exception is for the barren land cover class, which is the category with the fewest grid cells (1564 or 1.4% of the domain) and the land class with the highest average elevation (3178 m vs. a domain average of 2639 m). Comparing with SNODAS, Noah-MP-Def has smaller RMSE than Noah-MP-Cal. In all classes except cropland, Noah-MP-Cal has a lower RMSE when compared to UA.

Modeling in forested regions is often challenging due to uncertainty in snow-canopy interactions (Essery et al., 2009; Krinner et al., 2018). Therefore, we further subdivide the forest class into elevation bands to single out the impact of elevation on a land cover class with higher uncertainty (Figure 8b,d and see Table S4 for number of grid cells within each category). Results are similar to the full land cover comparison, where Noah-MP biases are negative, and Noah-MP-Cal has smaller bias and RMSE than Noah-MP-Def. Higher elevations have larger biases and RMSEs. At forested elevations below 3000 m, Noah-MP-Cal and Noah-MP-Def have similar values of RMSE. Calibration decreases errors in the higher elevation grid cells, which is often where more snow accumulates due to colder temperatures coupled with orographic lifting. We also calculate the ratio of RMSE to mean snow depth (not shown), and for Noah-MP-Cal, this metric decreases with elevation, while for Noah-MP-Def, it increases above 2500 m. Much of the increase in Noah-MP-Cal RMSE is due to deeper snowpacks at higher elevation. SNODAS and UA are both based on observational datasets, which likely have larger uncertainty in forests. Noah-MP-Cal is in better agreement with the observation-based gridded data products than Noah-MP-Def, but the "true" accuracy in forested environments is limited by a lack of observations in forests.

4.2.2 GHCN Comparisons

Across 79 GHCN stations, Noah-MP-Cal is less biased (0.0049 m) and has a lower RMSE (0.15 m) than Noah-MP-Def (bias of -0.04 m and RMSE of 0.20 m). Noah-MP-Cal generally reduces the snow depth bias in Noah-MP-Def in the Front Range and broadly reduces RMSE across the full domain (Figure 9). While results are generally similar between Noah-MP-Cal and Noah-MP-Def, the evaluation with GHCN demonstrates an additional independent check that calibration improves the performance of modeled snow depth.

4.2.3 SnowEx Comparisons

Finally, we also evaluate snow depth and SWE against 264 snow pit observations from the NASA SnowEx 2017 field campaigns at Grand Mesa and Senator Beck (Figure 10). Here we include SnowModel simulations in the comparison to consider a snow process model. For both Noah-MP and SnowModel, we select the grid that contains each snow pit for the comparison. SnowModel is kept at its native 30 m resolution, though we also tested average SnowModel grid cells to the Noah-MP resolution and results were similar. The majority of pit observations (n = 224) are from Grand Mesa, where there is better agreement after calibration for snow depth (Table 5): mean bias decreases from -48.2 cm to -12.1 cm (mean percent absolute difference decreases from 32.2% to 20.0%) and RMSE decreases from 54.4 cm to 34.9 cm. Similar for SWE, Noah-MP-Cal has a smaller SWE mean bias at Grand Mesa than Noah-MP-Def (-23.0 mm vs. -160.6 mm) and a smaller RMSE (132.9 mm vs 185.4 mm). For SWE, SnowModel has better agreement with snow pits than either Noah-MP simulation, though the performance of SnowModel and Noah-MP-Cal are comparable for snow depth, with Noah-MP-Cal having smaller MAE and RMSE. Similar performance for snow depth and SWE disagreements may be due to different density estimates in SnowModel and Noah-MP-Cal. At Senator Beck (n = 40 pits), where we do not have SnowModel simulations, Noah-MP-Cal greatly improves upon Noah-MP-Def evaluation metrics for both snow depth and SWE: for snow depth (SWE), RMSE increases from 49.7 cm to 102.5 cm (167.6 mm to 413.0). This highlights the uneven performance across the domain after calibration.

Spatially, Noah-MP-Def has much lower values of snow depth than measured in the snow pits on a single day (Figure 11). Noah-MP-Cal, on the other hand, has spatial patterns that better match the snow pits observations throughout the Grand Mesa study site, capturing the overall east-west gradient seen in the snow pit observations and in the SnowModel simulation. Calibrated Noah-MP at a 1-km resolution has similar error metrics to an uncalibrated snow process model at a 30-m resolution, but this evaluation is only possible over a small portion of the full domain.

Finally, we evaluate Noah-MP simulations against ASO lidar snow depth observations from SnowEx flights on February 8 and 16 (Painter, 2018). Spatially, estimates from ASO, Noah-MP-Cal, and Noah-MP-Def have somewhat similar patterns on each flight day, with snow depth tending to increase toward the eastern portion of the domain (Figure 12). ASO and Noah-MP-Cal also show that snow depth increases from the north to the south across the domain; Noah-MP-Def, on the other hand, has lower variability across the domain. Note the deeper band of snow in the ASO observations along the northern portion of the domain. The deeper snow here is likely due to snow accumulating at the base of the cliff. Snow persistence maps (Figure S3) show that snow historically lingers longer along the base of the cliff, suggesting that the deeper snow depths in ASO are plausible. Noah-MP, with grid cells orders of magnitude coarser than ASO, cannot capture this fine scale spatial pattern. For both flight days, Noah-MP-Cal has higher values of SPAEF, which indicates better spatial agreement with ASO observations: on February 8, Noah-MP-Cal has a SPAEF value of 0.408 and Noah-MP-Def has a value of

0.253; on February 16, Noah-MP-Cal has a SPAEF of 0.516 compared to 0.195 from Noah-MP-Def.

Originally collected at 3 m spatial resolution, ASO snow depth observations are aggregated to 0.01° resolution to match the Noah-MP simulations by averaging together over 100,000 ASO 3 m grid cells. In evaluations of Noah-MP grid cells against aggregated ASO depth observation, Noah-MP-Def underestimates ASO, and Noah-MP-Cal overestimates for snow depths above 1.5 m (Figure 12). For each flight day, Noah-MP-Cal has smaller RMSE, MAE, and bias magnitude than Noah-MP-Def. From this comparison, calibration may lead to overestimates of snow depth in some regions, but calibration introduces more realistic spatial patterns of snow depth, as compared to ASO observations.

4.3. Streamflow Evaluation

Beyond impacts on snow depth and SWE, calibration will impact LSM simulation of other hydrological variables. For six basins within the Colorado domain with little-to-no human management, calibration can improve streamflow estimation (Figure 13). Of the six basins, four have higher NSE values for Noah-MP-Cal than Noah-MP-Def. After calibration, however, four of the six basins still have negative NSE values, though the streamflow bias may not all be due to snow. For the two basins with NSE>0 (9072500 and 9081600), calibration improves performance, though only 9072500 has a statistically significant difference in monthly streamflow between Noah-MP-Cal and Noah-MP-Def. In two basins, both on the Gunnison River (9124700 and 9127800), calibration leads to a larger overestimation in streamflow for some evaluation years. For the Colorado River at Glenwood Springs (9072500), Noah-MP-Def largely underestimates streamflow, and calibration addresses this bias through increased runoff. In most years for most basins, Noah-MP-Cal has later peak streamflow, in agreement with the observations, which is also noted in Figure 13. In 9107000, where Noah-MP-Def overestimates observations, Noah-MP-Cal decreases the magnitude of the bias, though Noah-MP-Cal still overestimates slightly; in 9081600, Noah-MP-Def underestimates observed streamflow, and the calibrated runoff value is a better match for the observations. This demonstrates that calibration does not increase snow and runoff in one direction, but rather calibration can improve upon both positive and negative biases. Results similar to the small basin analysis are seen across the full model domain, including higher springtime streamflow in Noah-MP-Cal compared to Noah-MP-Def (Figure S4a,b,c). Peak streamflow in Noah-MP-Cal also generally occurs later in the year than Noah-MP-Def (Figure S4f), in agreement with later snowmelt in Noah-MP-Cal (Figures 3 and 8).

5. Discussion

5.1. Summary of results

Here we investigate the impact of model calibration on simulations of snow depth, SWE, and streamflow. From this calibration exercise, we aim to answer the three research questions posed in the introduction. First, calibration can address dry biases in LSMs, which often result in underestimation of snow. We show improvements to not only simulated SWE and snow depth magnitude but also to timing, of both accumulation and ablation periods. Calibration also results in Noah-MP-Cal performing about as well as an uncalibrated, high-resolution snow process model (Table 5, Figure 11). Though evaluations of Noah-MP and SnowModel are limited to Grand Mesa, the spatial variability of snow depth across the mesa are similar in SnowModel and Noah-MP-Cal, though SnowModel simulations produce more detail with the finer spatial resolution. When comparing both models to snow pit measurements, Noah-MP-Cal actually has better performance for snow depth, though error Noah-MP-Cal metrics are larger for SWE. Results are similar for high resolution ASO lidar, where Noah-MP-Cal captures the realistic spatial variability in ASO estimates, suggesting that, over Grand Mesa at least, the calibration procedure largely improves the model simulation.

Second, impacts from calibration are observed beyond snow variables. For streamflow, Noah-MP-Cal improves estimates for four of the six study basins. Here we are limited to small unmanaged basins or reconstructed estimates of natural streamflow since Noah-MP does not include human management. We note that we are not evaluating routed streamflow here, but instead, we consider grid cell estimates of surface and subsurface runoff. Future work should consider dynamically routed streamflow in order to account for time lags between the upper reaches of the watershed and the evaluation point with the stream gage. Even with those considerations, improved NSE metrics suggest that the increased snowpack in Noah-MP-Cal results in streamflow magnitude and timing that better matches observations.

Finally, the calibration highlights potential avenues for improving both model configuration and meteorological forcing data, though calibrated parameters may be reflective of the choice of forcing dataset (Elsner et al., 2014). The genetic algorithm procedure produces spatially varying model parameters, as compared to the spatially uniform parameters used in the default Noah-MP configuration. In particular, we highlight ten parameters within Noah-MP that are likely candidates for further investigation. We show that allowing these parameters to vary in space results in improved model performance compared to the default, spatially uniform values. Some of the parameters, such as $SNOWF\_SCALE\_$, appear to have a relationship with elevation (compare Figure 2f with Figure 1), while other parameters, such as $MXSNALB$ and $Z0SNO$, appear to be more related to land class category (compare Figure 2b,j with Figure 1). Future efforts should determine new estimates for these parameters, perhaps through investigation of relationships with landscape characteristics, such as elevation, vegetation class, and soil type.

Global maps of the sensitive parameters could likely improve simulation of snow without the need for a computationally expensive calibration procedure. In addition to investigating relationships for creating spatially varying parameters, work should consider whether parameters should also vary in time. Creating new estimates of spatially and temporally varying parameters could improve snow modeling without the data requirement of calibration, which would have implications for our ability to estimate global snow, regardless of data availability. Efforts to scale snow parameters examined here to larger domains are under development, resulting in spatially varying parameter estimates for all of CONUS.

In addition to the ten parameters from Noah-MP discussed above, results from calibration demonstrate that introducing the snowfall scaling term has a large impact on the snow accumulation magnitude. This points to the need for better meteorological forcing data, particularly for precipitation at high elevations. There is often high variability between precipitation estimates from differing models and reanalyses (Decker et al., 2012; Essou et al., 2016; Henn et al., 2018; Hughes et al., 2017; Wrzesien et al., 2019), and previous studies have suggested that NLDAS-2 precipitation is too low in mountain regions (Enzminger et al., 2019; He et al., 2019; Henn et al., 2018; Smyth et al., 2020); such uncertainty will be propagated into the LSM. However, improving large scale precipitation estimates is not trivial, and model-based precipitation estimates often outperform observation-based estimates in mountain areas (Lundquist et al., 2019), despite known model biases. If we cannot improve estimates of precipitation and snowfall in the forcing datasets, informing modeled snowpack estimates with observations of SWE and snow depth is likely the best option. This calibration procedure highlights a method for addressing biases in both meteorological forcing and the LSM itself and results in improved simulations of snow in a topographically complex region.

5.2. Implications for Snow OSSE

As discussed, the Noah-MP-Cal simulation presented here will be used as the nature run (NR) in a snow-focused Observing System Simulation Experiment (OSSE), where the NR is designed to approximate the "truth", i.e., actual snow conditions. Though calibration is not a panacea for reducing all model uncertainty, the improved performance from Noah-MP-Def to Noah-MP-Cal provides compelling support for Noah-MP-Cal to be the NR for the OSSE. Of particular concern when designing the OSSE was whether the NR could address the common underestimation of snow at higher elevations, which is necessary for understanding how proposed sensors will observe realistic ranges of snow conditions. Calibrating Noah-MP against UA SWE estimates reduces the negative bias for SWE and snow depth and results in snow spatial heterogeneity that better matches both UA and SNODAS.

While Noah-MP-Cal is not without error, a NR is not expected to perfectly replicate actual conditions, and no true observations are used in an OSSE. Therefore, the spatial and temporal variability in Noah-MP-Cal is adequate for approximating realistic snow conditions for the western Colorado domain. The main drawback of Noah-MP as the NR – whether the default or calibrated configuration – is that Noah-MP does not provide estimates of snow grain size. Understanding how satellite observations are impacted by snow grain size and metamorphism is a fundamentally important question (Durand et al, 2018; Nolin, 2010; Foster et al., 2005). However, no models within the current LIS framework provide estimates of snow grain size, though work is ongoing to implement new snow models into LIS. While Noah-MP-Cal will be used in the OSSE described here, future work will consider a follow on OSSE that incorporates a model that does include the simulation of grain size.

5.3. Challenges with Calibration

With a calibration exercise such as this one, there are a few notable challenges. While calibration can lead to domain-averaged improvements in the targeted variable, as presented here for SWE and snow depth, it can cause degraded performance in individual regions across the domain. We see this in the northern portion of the domain (Figure 4a) to the west of the Cameron Pass evaluation site (Figure 1). After the genetic algorithm optimization, the snowfall scale term is high in this region (Figure 2k), resulting in snow depths and SWE values that much larger than either UA or SNODAS. In the calibration period, SWE estimates from UA were particularly large, where the 2007-2009 average peak SWE value for this area from UA is higher than the average peak SWE value for 2010-2020, causing the calibration to be trained on higher-than average SWE. Anomalies such as this from calibration are often unavoidable.

Another challenge with our calibration setup is that the parameters are constant in time. Therefore, even if Noah-MP-Def performs well compared to UA, the calibrated parameters will still be applied. For example, in water years 2017, 2019, and 2020, domain-averaged SWE and snow depth from Noah-MP-Def is similar to UA (Figure 3). Applying the calibrated parameters generally results in increased snow values, and as a result, Noah-MP-Cal overestimates SWE in these years. Calibration improves performance over the full study period (Table 3), but it does not always result in better performance for an individual year or season. As discussed above with the spatial anomalies, calibration will not result in uniformly improved performance.

For all calibration procedures, such as the genetic algorithm used here, a “truth” dataset is required to calibrate against, and data availability is limited in many regions, especially in high elevations and high latitudes where much of the global snow accumulates. Therefore, while the calibration procedure presented here is a critical step for the ongoing OSSE and for improving the representation of the truth, calibrating over a well-observed Colorado domain may not necessarily improve the model performance of global snow. Results presented here may not reflect other regions with differing snow conditions, such as maritime snow in the Pacific Northwest or tundra snow in the high latitudes (e.g., Kim et al., 2021). Future work will investigate similar calibration methods in other regions. We hypothesize that in regions with high precipitation uncertainty, such as mountainous regions, the snowfall scaling term will have similar impacts on snow magnitude as presented here.

Since we only calibrate against SWE and do not include additional constraints in the objective function, such as for streamflow, the calibration cannot directly address biases in other model processes. In the streamflow analyses, we see that Noah-MP-Def does not have good agreement with the observed runoff (Figure 13 and Table 6). However, further observational constraints or model improvements (possibly unrelated to snow processes) are required to address runoff biases that we show here. In operational modeling, it is standard to calibrate snowmelt rates to runoff (e.g. Hay et al., 2006; Franz and Karsten, 2013; Turcotte et al., 2017), in order to constrain snow ablation. Here, though, we do not calibrate against runoff. Degradation in unconstrained variables, such as runoff, are not uncommon during calibration efforts (e.g. Franz and Karsten, 2013; Nemri and Kinnard). Future efforts could consider multi-criteria objective functions to reduce biases in both snow variables and streamflow.

Beyond the calibration, evaluating gridded data with point observations presents additional challenges. There are significant differences in what an ~1 m observation, such as a snow pit or a GHCN station, measures and what a ~1000 m model grid cell simulates. Since snow depth and SWE measurements are typically point observations, this imperfect comparison is often necessary for evaluating models. However, during extensive field campaigns, such as SnowEx 2017, numerous observations are made in a small domain over a short period of time. While the result is still point-to-grid comparisons, the high density of observation allows for a more complete evaluation, if over a limited domain. Though we do acknowledge the uncertainty from scale differences, we aim to provide as thorough an evaluation as we can for demonstrating evidence of the improved performance of Noah-MP-Cal over Noah-MP-Def through the numerous independent comparison datasets.

6. Conclusion

The Noah-MP-Cal and Noah-MP-Def evaluation demonstrates that calibrating a land surface model against an observation-based SWE dataset (e.g., the University of Arizona dataset), improves model performance of snow, though not uniformly across the domain. The calibration procedure was motivated by an ongoing Observing System Simulation Experiment (OSSE) to evaluate the utility of proposed snow satellite sensors, and Noah-MP-Cal will be used as the nature run for the OSSE. That is, the improved Noah-MP simulation will act as the “truth” for the OSSE, upon which synthetic observations will be created. However, results presented here have important implications beyond the OSSE. We demonstrate a method for improving spatiotemporal estimates of snow, and we show that spatially uniform values of key model parameters result in worse performance. Allowing parameters to vary spatially, as we do in the Noah-MP-Cal simulation after a genetic algorithm optimization procedure, results in improved model performance of both snow depth and SWE. Future model development could consider implementing distributed values of sensitive parameters, which might improve LSM simulations without the need for an initial calibration step.

7. Acknowledgements

This research was supported by Grants from the National Aeronautics and Space Administration (#AIST18-0041, #AIST18-0045, and #NNH16ZDA001N). Computing was supported by the resources at the NASA Center for Climate Simulations. We thank three anonymous reviewers for feedback that strengthened the overall manuscript.

8. Data Availability Statement

All datasets described here for modeling forcing and evaluation are available through the provided citations within the text. The University of Arizona data are available for download from the National Snow and Ice Data Center (NSIDC, Broxton et al., 2019). The Noah-MP model simulations upon which this study is based are too large to be publicly archived with available resources, though all model output are stored on the NASA Discover supercomputer system through the NASA Center for Climate Simulations. To replicate the simulation, interested users can access the NASA Land Information System at https://github.com/NASA-LIS/LISF.

9. References

Aguado, E. (1985). Radiation Balances of Melting Snow Covers at an Open Site in the Central Sierra Nevada, California. *Water Resources Research*, *21*(11), 1649–1654. https://doi.org/10.1029/WR021i011p01649

Ahl, R. S., Woods, S. W., & Zuuring, H. R. (2008). Hydrologic Calibration and Validation of SWAT in a Snow-Dominated Rocky Mountain Watershed, Montana, U.S.A.1. *JAWRA Journal of the American Water Resources Association*, *44*(6), 1411–1430. https://doi.org/10.1111/j.1752-1688.2008.00233.x

Amorocho, J., & Espildora, B. (1966). *Mathematical Simulation of the Snow Melting Process*. Department of Water Science and Engineering, University of California, Davis.

Anderson, E. (1973), National Weather Service River Forecast system—Snow accumulation and ablation model, NOAA Tech. Memo. NWS HYDRO-17, 217 pp., U.S. Dep. of Commer., Silver Spring, Md.

Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, *438*(7066), 303–309. https://doi.org/10.1038/nature04141

Barrett, A. P. (2003), National Operational Hydrologic Remote Sensing Center Snow Data Assimilation System (SNODAS) Products at NSIDC, 19 pp., Natl. Snow and Ice Data Cent., Coop. Inst. for Res. in Environ. Sci., Boulder, Colo.

Bormann, K. J., Brown, R. D., Derksen, C., & Painter, T. H. (2018). Estimating snow-cover trends from space. *Nature Climate Change*, *8*(11), 924–928. https://doi.org/10.1038/s41558-018-0318-3

Broxton, P. D., Dawson, N., & Zeng, X. (2016). Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth. *Earth and Space Science*, *3*(6), 246–256. https://doi.org/10.1002/2016EA000174

Broxton, P. D., Zeng, X., & Dawson, N. (2016). Why Do Global Reanalyses and Land Data Assimilation Products Underestimate Snow Water Equivalent? *Journal of Hydrometeorology*, *17*(11), 2743–2761. https://doi.org/10.1175/JHM-D-16-0056.1

Broxton, P., X. Zeng, and N. Dawson. 2019. *Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1*. [Indicate subset used]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/0GGPB220EX6A.

Carroll, T., Cline, D., Fall, G., Nilsson, A., Li, L., & Rost, A. 2001. NOHRSC operations and the simulation of snow cover properties for the coterminous US. In Proc. 69th Annual Meeting of the Western Snow Conf (pp. 1-14).

Chen, Fei, Barlage, M., Tewari, M., Rasmussen, R., Jin, J., Lettenmaier, D., et al. (2014). Modeling seasonal snowpack evolution in the complex terrain and forested Colorado Headwaters region: A model intercomparison study. *Journal of Geophysical Research: Atmospheres*, *119*(24), 13,795-13,819. https://doi.org/10.1002/2014JD022167

Chen, Feng, Liu, C., Dudhia, J., & Chen, M. (2014). A sensitivity study of high-resolution regional climate simulations to three land surface models over the western United States. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7271–7291. https://doi.org/10.1002/2014JD021827

Cho, E., & Jacobs, J. M. (2020). Extreme value snow water equivalent and snowmelt for infrastructure design over the contiguous United States. *Water Resources Research*, 56, e2020WR028126. https://doi.org/10.1029/2020WR028126

Cho, E., Jacobs, J. M., & Vuyovich, C. M. (2020). The Value of Long-Term (40 years) Airborne Gamma Radiation SWE Record for Evaluating Three Observation-Based Gridded SWE Data Sets by Seasonal Snow and Land Cover Classifications. *Water Resources Research*, *56*(1), e2019WR025813. https://doi.org/10.1029/2019WR025813

Chow, V.T. (1964) Handbook of Applied Hydrology. McGraw-Hill Book Company, New York.

Clow, D. W., Nanus, L., Verdin, K. L., & Schmidt, J. (2012). Evaluation of SNODAS snow depth and snow water equivalent estimates for the Colorado Rocky Mountains, USA. *Hydrological Processes*, *26*(17), 2583–2591. https://doi.org/10.1002/hyp.9385

Crow, W. T., Drusch, M., & Wood, E. F. (2001). An observation system simulation experiment for the impact of land surface heterogeneity on AMSR-E soil moisture retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8), 1622-1631, https://doi.org/10.1109/36.942540.

Crow, W. T. *et al*. (2005). An observing system simulation experiment for hydros radiometer-only soil moisture products. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6), 1289-1303. https://doi.org/10.1109/TGRS.2005.845645.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., et al. (2016). The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. *Journal of Geophysical Research: Atmospheres*, *121*(18), 10,676-10,700. https://doi.org/10.1002/2016JD025097

Daly, C., G. H. Taylor, W. P. Gibson, T. W. Parzybok, G. L. Johnson, & P. A. Pasteris. (2000). HIGH-QUALITY SPATIAL CLIMATE DATA SETS FOR THE UNITED STATES AND BEYOND. *Transactions of the ASAE*, *43*(6), 1957–1962. https://doi.org/10.13031/2013.3101

Dawson, N., Broxton, P., & Zeng, X. (2017). A New Snow Density Parameterization for Land Data Initialization. *Journal of Hydrometeorology*, *18*(1), 197–207. https://doi.org/10.1175/JHM-D-16-0166.1

Dawson, N., Broxton, P., & Zeng, X. (2018). Evaluation of Remotely Sensed Snow Water Equivalent and Snow Cover Extent over the Contiguous United States. *Journal of Hydrometeorology*, *19*(11), 1777–1791. https://doi.org/10.1175/JHM-D-18-0007.1

Decker, M., Brunke, M. A., Wang, Z., Sakaguchi, K., Zeng, X., & Bosilovich, M. G. (2012). Evaluation of the Reanalysis Products from GSFC, NCEP, and ECMWF Using Flux Tower Observations. *Journal of Climate*, *25*(6), 1916–1944. https://doi.org/10.1175/JCLI-D-11-00004.1

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2), 1299-1315. https://doi.org/10.5194/hess-22-1299-2018

Dirmhirn, I., & Eaton, F. D. (1975). Some Characteristics of the Albedo of Snow. *Journal of Applied Meteorology*, *14*(3), 375–379. https://doi.org/10.1175/1520-0450(1975)014<0375:SCOTAO>2.0.CO;2

Dozier, J., Bair, E. H., & Davis, R. E. (2016). Estimating the spatial distribution of snow water equivalent in the world's mountains: Spatial distribution of snow in the mountains. *Wiley Interdisciplinary Reviews: Water*, *3*(3), 461–474. https://doi.org/10.1002/wat2.1140

Duethmann, D., Peters, J., Blume, T., Vorogushyn, S., & Güntner, A. (2014). The value of satellite-derived snow cover images for calibrating a hydrological model in snow-dominated catchments in Central Asia. *Water Resources Research*, *50*(3), 2002–2021. https://doi.org/10.1002/2013WR014382

Durand, M., Gatebe, C., Kim, E., Molotch, N., Painter, T.H., Raleigh, M., Sandells, M. & Vuyovich, C. (2018). NASA SnowEx Science Plan: Assessing approaches for measuring water in Earth's Seasonal Snow, version 1.6.

Dutra, E., Viterbo, P., Miranda, P. M. A., & Balsamo, G. (2012). Complexity of Snow Schemes in a Climate Model and Its Impact on Surface Energy and Hydrology. *Journal of Hydrometeorology*, *13*(2), 521–538. https://doi.org/10.1175/JHM-D-11-072.1

Elder, K., L. Brucker, C. Hiemstra, and H. Marshall. 2018. *SnowEx17 Community Snow Pit Measurements, Version 1*. [Indicate subset used]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: https://doi.org/10.5067/Q0310G1XULZS.

Elsner, M. M., Gangopadhyay, S., Pruitt, T., Brekke, L. D., Mizukami, N., & Clark, M. P. (2014). How Does the Choice of Distributed Meteorological Data Affect Hydrologic Model Calibration and Streamflow Simulations? *Journal of Hydrometeorology*, *15*(4), 1384–1403. https://doi.org/10.1175/JHM-D-13-083.1

Enzminger, T. L., Small, E. E., & Borsa, A. A. (2019). Subsurface Water Dominates Sierra Nevada Seasonal Hydrologic Storage. *Geophysical Research Letters*, *46*(21), 11993–12001. https://doi.org/10.1029/2019GL084589

Essery, R., Rutter, N., Pomeroy, J., Baxter, R., Stähli, M., Gustafsson, D., et al. (2009). SNOWMIP2: An Evaluation of Forest Snow Process Simulations. *Bulletin of the American Meteorological Society*, *90*(8), 1120–1136. https://doi.org/10.1175/2009BAMS2629.1

Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F., & Poulin, A. (2016). Can Precipitation and Temperature from Meteorological Reanalyses Be Used for Hydrological Modeling? *Journal of Hydrometeorology*, *17*(7), 1929–1950. https://doi.org/10.1175/JHM-D-15-0138.1

Etchevers, P., Martin, E., Brown, R., Fierz, C., Lejeune, Y., Bazile, E., et al. (2004). Validation of the energy budget of an alpine snowpack simulated by several snow models

(Snow MIP project). *Annals of Glaciology*, *38*, 150–158. https://doi.org/10.3189/172756404781814825

Foster, J. L., Sun, C., Walker, J. P., Kelly, R., Chang, A., Dong, J., & Powell, H. (2005). Quantifying the uncertainty in passive microwave snow water equivalent observations. *Remote Sensing of environment*, *94*(2), 187-203.

Franz, K. J., Hogue, T. S., & Sorooshian, S. (2008). Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts. *Journal of Hydrology*, *360*(1), 48–66. https://doi.org/10.1016/j.jhydrol.2008.07.013

Franz, K. J., Butcher, P., & Ajami, N. K. (2010). Addressing snow model uncertainty for hydrologic prediction. *Advances in Water Resources*, *33*(8), 820–832. https://doi.org/10.1016/j.advwatres.2010.05.004

Franz, K. J., & Karsten, L. R. (2013). Calibration of a distributed snow model using MODIS snow covered area data. *Journal of Hydrology*, *494*, 160–175. https://doi.org/10.1016/j.jhydrol.2013.04.026

Garnaud, C., Bélair, S., Carrera, M. L., Derksen, C., Bilodeau, B., Abrahamowicz, M., Gauthier, N., & Vionnet, V. (2019). Quantifying Snow Mass Mission Concept Trade-Offs Using an Observing System Simulation Experiment, *Journal of Hydrometeorology*, *20*(1), 155-173.

Hall, D. K. and G. A. Riggs. 2016. *MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, Version 6*. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A1.006.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., & Bayr, K. J. (2002). MODIS snow-cover products. *Remote Sensing of Environment*, *83*(1–2), 181–194. https://doi.org/10.1016/S0034-4257(02)00095-0

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850-853. https://doi.org/10.1126/science.1244693

Harrison, K. W., Kumar, S. V., Peters-Lidard, C. D., and Santanello, J. A. (2012), Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques, Water Resour. Res., 48, W11514, doi:10.1029/2012WR012337.

He, C., Chen, F., Barlage, M., Liu, C., Newman, A., Tang, W., et al. (2019). Can Convection-Permitting Modeling Provide Decent Precipitation for Offline High-Resolution Snowpack Simulations Over Mountains? *Journal of Geophysical Research: Atmospheres*, *124*(23), 12631–12654. https://doi.org/10.1029/2019JD030823

He, M., Hogue, T. S., Franz, K. J., Margulis, S. A., & Vrugt, J. A. (2011). Characterizing parameter sensitivity and uncertainty for a snow model across hydroclimatic regimes. Advances in Water Resources, 34(1), 114–127. https://doi.org/10.1016/j.advwatres.2010.10.002

Henn, B., Clark, M. P., Kavetski, D., McGurk, B., Painter, T. H., & Lundquist, J. D. (2016). Combining snow, streamflow, and precipitation gauge observations to infer basin-mean precipitation. *Water Resources Research*, *52*(11), 8700–8723. https://doi.org/10.1002/2015WR018564

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Holtzman, N. M., Pavelsky, T. M., Cohen, J. S., Wrzesien, M. L., & Herman, J. D. (2020). Tailoring WRF and Noah-MP to Improve Process Representation of Sierra Nevada Runoff: Diagnostic Evaluation and Applications. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001832. https://doi.org/10.1029/2019MS001832

Hughes, M., Lundquist, J. D., & Henn, B. (2017). Dynamical downscaling improves upon gridded precipitation products in the Sierra Nevada, California. *Climate Dynamics*. https://doi.org/10.1007/s00382-017-3631-z

Immerzeel, W.W., Lutz, A.F., Andrade, M. *et al.* Importance and vulnerability of the world's water towers. *Nature* **577,** 364–369 (2020). https://doi.org/10.1038/s41586-019-1822-y

Isenstein, E. M., Wi, S., & Ethan Yang, Y. C. (2015). Calibration of a Distributed Hydrologic Model Using Streamflow and Remote Sensing Snow Data. In *World Environmental and Water Resources Congress 2015* (pp. 973–982). Austin, TX: American Society of Civil Engineers. https://doi.org/10.1061/9780784479162.093

Jordan, R. (1991). A one-dimensional temperature model for a snow cover: technical documentation for SNTHERM.89. CRREL Spec. Rep. 91-16.

Kim, R. S., Kumar, S., Vuyovich, C., Houser, P., Lundquist, J., Mudryk, L., et al. (2021). Snow Ensemble Uncertainty Project (SEUP): quantification of snow water equivalent uncertainty across North America via ensemble land surface modeling, *The Cryosphere*, 15, 771–791, https://doi.org/10.5194/tc-15-771-2021

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873-1886. https://doi.org/10.5194/gmd-11-1873-2018

Koster, R. D., Mahanama, S. P., Livneh, B., Lettenmaier, D. P., & Reichle, R. H. (2010). Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nature Geoscience*, *3*(9), 613-616.Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., et al. (2018). ESM-SnowMIP: assessing snow models and quantifying snow-related climate feedbacks. *Geoscientific Model Development*, *11*(12), 5027–5049. https://doi.org/10.5194/gmd-11-5027-2018

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., et al. (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software*, *21*(10), 1402–1415. https://doi.org/10.1016/j.envsoft.2005.07.004

Kumar, S. V., Peters-Lidard, C. D., Santanello, J., Harrison, K., Liu, Y., & Shaw, M. (2012). Land surface Verification Toolkit (LVT) – a generalized framework for land surface model evaluation. *Geoscientific Model Development*, *5*(3), 869–886. https://doi.org/10.5194/gmd-5-869-2012

Kumar, Sujay V., M. Mocko, D., Wang, S., Peters-Lidard, C. D., & Borak, J. (2019). Assimilation of Remotely Sensed Leaf Area Index into the Noah-MP Land Surface Model: Impacts on Water and Carbon Fluxes and States over the Continental United States. *Journal of Hydrometeorology*, *20*(7), 1359–1377. https://doi.org/10.1175/JHM-D-18-0237.1

Kumar, S. V., Reichle, R. H., Harrison, K. W., Peters-Lidard, C. D., Yatheendradas, S., & Santanello, J. A. (2012). A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, *48*(3). https://doi.org/10.1029/2010WR010261

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., & Wood, E. F. (2015). Inroads of remote sensing into hydrologic science during the WRR era: REMOTE SENSING. *Water Resources Research*, *51*(9), 7309–7342. https://doi.org/10.1002/2015WR017616

Li, D., Wrzesien, M. L., Durand, M., Adam, J., & Lettenmaier, D. P. (2017). How much runoff originates as snow in the western United States, and how will that change in the future?: Western U.S. Snowmelt-Derived Runoff. *Geophysical Research Letters*, *44*(12), 6163–6172. https://doi.org/10.1002/2017GL073551

Liston, G. E., & Elder, K. (2006). A distributed snow-evolution modeling system (SnowModel). Journal of Hydrometeorology, 7(6), 1259–1276. https://doi.org/10.1175/JHM548.1

Liston, G. E., & Elder, K. (2006b). A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). Journal of Hydrometeorology, 7(2), 217–234. https://doi.org/10.1175/JHM486.1

Liston, G. E., Haehnel, R. B., Sturm, M., Hiemstra, C. A., Berezovskaya, S., & Tabler, R. D. (2007). Simulating complex snow distributions in windy environments using SnowTran-3D. Journal of Glaciology, 53(181), 241–256. https://doi.org/10.3189/172756507782202865

Liston, G. E., & Sturm, M. (1998). A snow-transport model for complex terrain. Journal of Glaciology, 44(148), 498–516. https://doi.org/10.1017/S0022143000002021

Lundquist, J., Hughes, M., Gutmann, E., & Kapnick, S. (2019). Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bulletin of the American Meteorological Society*, BAMS-D-19-0001.1. https://doi.org/10.1175/BAMS-D-19-0001.1

Magnusson, J., Winstral, A., Stordal, A. S., Essery, R., & Jonas, T. (2017). Improving physically based snow simulations by assimilating snow depths using the particle filter. Water Resources Research, 53(2), 1125–1143. https://doi.org/10.1002/2016WR019092

Marks, D., & Dozier, J. (1992). Climate and energy exchange at the snow surface in the Alpine Region of the Sierra Nevada: 2. Snow cover energy balance. *Water Resources Research*, *28*(11), 3043–3054. https://doi.org/10.1029/92WR01483

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, *51*(1), 716–728. https://doi.org/10.1002/2014WR015820

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Minder, J. R., Letcher, T. W., & Skiles, S. M. (2016). An evaluation of high-resolution regional climate model simulations of snow cover and albedo over the Rocky Mountains, with implications for the simulated snow-albedo feedback: Evaluating mountain snow cover in RCMs. *Journal of Geophysical Research: Atmospheres*, *121*(15), 9069–9088. https://doi.org/10.1002/2016JD024995

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G. S., Crow, W. T., Thorp, K. R., Moran, M. S., Reichle, R. H., and Gupta, H. V. (2012), Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment, *Water Resources Research*, 48, W05525. https://doi.org/10.1029/2011WR011420.

Nemri, S., & Kinnard, C. (2020). Comparing calibration strategies of a conceptual snow hydrology model and their impact on model performance and parameter identifiability. *Journal of Hydrology*, *582*, 124474.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A., K. Sampson, M.P. Clark, A. Bock, and R.J. Viger, and D. Blodgett, 2014: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR. doi:10.5065/D6MW2F4D

Niu, G.-Y., & Yang, Z.-L. (2007). An observation-based formulation of snow cover fraction and its evaluation over large North American river basins. *Journal of Geophysical Research*, *112*(D21), D21101. https://doi.org/10.1029/2007JD008674

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*(D12), D12109. https://doi.org/10.1029/2010JD015139

Nolin, A. W. (2010). Recent advances in remote sensing of seasonal snow. *Journal of Glaciology*, *56*(200), 1141–1150. https://doi.org/10.3189/002214311796406077

Painter, T. (2018). ASO L4 LiDAR snow depth 3m UTM grid, version 1. Grand Mesa Colorado. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/KIE9QNVG7HP0

Parajka, J., & Blöschl, G. (2008). The value of MODIS snow cover data in validating and calibrating conceptual hydrologic models. *Journal of Hydrology*, *358*(3), 240–258. https://doi.org/10.1016/j.jhydrol.2008.06.006

Parajka, J., Merz, R., & Blöschl, G. (2007). Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments. *Hydrological Processes*, *21*(4), 435–446. https://doi.org/10.1002/hyp.6253

Peters-Lidard, C. D., Houser, P. R., Tian, Y., Kumar, S. V., Geiger, J., Olden, S., et al. (2007). High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innovations in Systems and Software Engineering*, *3*(3), 157–165. https://doi.org/10.1007/s11334-007-0028-x

Pomeroy, J. W., Gray, D. M., Shook, K. R., Toth, B., Essery, R. L. H., Pietroniro, A., & Hedstrom, N. (1998). An evaluation of snow accumulation and ablation processes for land surface modelling. *Hydrological Processes*, *12*(15), 2339–2367. https://doi.org/10.1002/(SICI)1099-1085(199812)12:15<2339::AID-HYP800>3.0.CO;2-L

Raleigh, M. S., Lundquist, J. D., & Clark, M. P. (2015). Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. Hydrology and Earth System Sciences, 19(7), 3153–3179. https://doi.org/10.5194/hess-19-3153-2015

Reba, M. L., Marks, D., Link, T. E., Pomeroy, J., & Winstral, A. (2014). Sensitivity of model parameterizations for simulated latent heat flux at the snow surface for complex mountain sites. *Hydrological Processes*, *28*(3), 868–881. https://doi.org/10.1002/hyp.9619

Rittger, K., Painter, T. H., & Dozier, J. (2013). Assessment of methods for mapping snow cover from MODIS. *Advances in Water Resources*, *51*, 367–380. https://doi.org/10.1016/j.advwatres.2012.03.002

Rollins, M. G.: LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment, Int. J. Wildl. Fire, 18(3), 235–249. https://doi.org/10.1071/WF08088, 2009.

Rutter, N., Essery, R., Pomeroy, J., Altimir, N., Andreadis, K., Baker, I., et al. (2009). Evaluation of forest snow processes models (SnowMIP2). *Journal of Geophysical Research*, *114*(D6), D06111. https://doi.org/10.1029/2008JD011063

Schmucki, E., Marty, C., Fierz, C., & Lehning, M. (2014). Evaluation of modelled snow depth and snow water equivalent at three contrasting sites in Switzerland using SNOWPACK simulations driven by different meteorological data input. Cold Regions Science and Technology, 99, 27–37. https://doi.org/10.1016/j.coldregions.2013.12.004

Shafii, M., & De Smedt, F. (2009). Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrology and Earth System Sciences*, *13*(11), 2137–2149. https://doi.org/10.5194/hess-13-2137-2009

Skiles, S. M. (2018). Grand Mesa study plot (version 1), Zenodo. https://doi.org/10.5281/zenodo.1479859

Smyth, E. J., Raleigh, M. S., & Small, E. E. (2020). Improving SWE Estimation With Data Assimilation: The Influence of Snow Depth Observation Timing and Uncertainty. *Water Resources Research*, *56*(5), e2019WR026853. https://doi.org/10.1029/2019WR026853

Swain, M. J. and Ballard, D. H.: Color indexing, Int. J. Comput. Vis., 7, 11–32, https://doi.org/10.1007/BF00130487, 1991.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., et al. (2011). Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. *Remote Sensing of Environment*, *115*(12), 3517–3529. https://doi.org/10.1016/j.rse.2011.08.014

ter Braak, C. J., & Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435-446.

Viviroli, D., Dürr, H. H., Messerli, B., Meybeck, M. & Weingartner, R. Mountains of the world, water towers for humanity: typology, mapping, and global significance. *Water Resources Research,* **43**, 1–13 (2007). https://doi.org/10.1029/2006WR005653

Verseghy, D.L. (1991), Class—A Canadian land surface scheme for GCMS. I. Soil model. Int. J. Climatol., 11: 111-133. https://doi.org/10.1002/joc.3370110202

Vuyovich, C. M., Jacobs, J. M., & Daly, S. F. (2014). Comparison of passive microwave and modeled estimates of total watershed SWE in the continental United States. *Water Resources Research*, *50*(11), 9088–9102. https://doi.org/10.1002/2013WR014734

Wang, Q. J. (1991). The Genetic Algorithm and Its Application to Calibrating Conceptual Rainfall-Runoff Models. *Water Resources Research*, *27*(9), 2467–2471. https://doi.org/10.1029/91WR01305

Wang, X., Barker, D. M., Snyder, C., & Hamill, T. M. (2008). A Hybrid ETKF–3DVAR Data Assimilation Scheme for the WRF Model. Part I: Observing System Simulation Experiment, *Monthly Weather Review*, *136*(12), 5116-5131.

Wang, Y.-H., Broxton, P., Fang, Y., Behrangi, A., Barlage, M., Zeng, X., & Niu, G.-Y. (2019). A Wet-Bulb Temperature-Based Rain-Snow Partitioning Scheme Improves Snowpack Prediction Over the Drier Western United States. *Geophysical Research Letters*, *46*(23), 13825–13835. https://doi.org/10.1029/2019GL085722

Webb, R. W., Raleigh, M. S., McGrath, D., Molotch, N. P., Elder, K., & Hiemstra, C., et al. (2020). Within-stand boundary effects on snow water equivalent distribution in forested areas. Water Resources Research, 56, e2019WR024905. https://doi.org/10.1029/2019WR024905

Wrzesien, M. L., Pavelsky, T. M., Kapnick, S. B., Durand, M. T., & Painter, T. H. (2015). Evaluation of snow cover fraction for regional climate simulations in the Sierra Nevada. *International Journal of Climatology*, *35*(9), 2472–2484. https://doi.org/10.1002/joc.4136

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Howat, I. M., Margulis, S. A., & Huning, L. S. (2017). Comparison of Methods to Estimate Snow Water Equivalent at the Mountain Range Scale: A Case Study of the California Sierra Nevada. *Journal of Hydrometeorology*, *18*(4), 1101–1119. https://doi.org/10.1175/JHM-D-16-0246.1

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Kapnick, S. B., Zhang, Y., Guo, J., & Shum, C. K. (2018). A New Estimate of North American Mountain Snow Accumulation From Regional Climate Model Simulations. *Geophysical Research Letters*, *45*(3), 1423–1432. https://doi.org/10.1002/2017GL076664

Wrzesien, M. L., Durand, M. T., & Pavelsky, T. M. (2019). A Reassessment of North American River Basin Cool-Season Precipitation: Developments From a New Mountain Climatology Data Set. *Water Resources Research*, *55*(4), 3502–3519. https://doi.org/10.1029/2018WR024106

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products: water and energy flux analysis. *Journal of Geophysical Research: Atmospheres*, *117*(D3). https://doi.org/10.1029/2011JD016048

Xia, Y., Mocko, D., Huang, M., Li, B., Rodell, M., Mitchell, K. E., et al. (2017). Comparison and Assessment of Three Advanced Land Surface Models in Simulating Terrestrial Water Storage Components over the United States. *Journal of Hydrometeorology*, *18*(3), 625–649. https://doi.org/10.1175/JHM-D-16-0112.1

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, *204*(1), 83–97. https://doi.org/10.1016/S0022-1694(97)00107-8

Zeng, X., Broxton, P., & Dawson, N. (2018). Snowpack Change From 1982 to 2016 Over Conterminous United States. *Geophysical Research Letters*, *45*(23), 12,940-12,947. https://doi.org/10.1029/2018GL079621

10. Tables

Table 1. Calibration parameters including default values, calibration range, and average calibrated value. The calibration range reference, when applicable, is noted. Otherwise, the calibration range is +/- 20% of the default value.

Parameter	Description [units]	Default Value	Calibration Range	Calibrated Average Value	Reference
ALBDRY1	Dry soil albedos (VIS) [-]	0.10-0.27	0.08-0.32	0.200	--
ALBDRY2	Dry soil albedos (NIR) [-]	0.20-0.54	0.16-0.65	0.404	--
ALBICE1	Albedo land ice (VIS) [-]	0.55	0.44-0.66	0.552	--
ALBICE2	Albedo land ice (NIR) [-]	0.80	0.64-0.96	0.795	--
ALBSAT1	Saturated soil albedos (VIS) [-]	0.05-0.15	0.04-0.18	0.110	--
ALBSAT2	Saturated soil albedos (NIR) [-]	0.10-0.30	0.08-0.36	0.224	--
BETADS	Two stream parameter β_d for snow [-]	0.5	0.4-0.6	0.498	--
BETAIS	Two stream parameter βi for snow [-]	0.5	0.4-0.6	0.501	--
EG1	Emissivity soil surface (soil) [-]	0.97	0.78-1.0	0.895	--
EG2	Emissivity soil surface (lake) [-]	0.98	0.78-1.0	0.891	--
MFSNO	Snowmelt parameter [-]	2.5	0.5-3.0	1.158	Niu & Yang (2007)
MNSNALB	Minimum snow albedo [-]	0.55	0.45-0.65	0.596	Aguado (1985); Dirmhirn & Eaton (1975)

MXSNALB	Maximum snow albedo [-]	0.84	0.75-0.95	0.853	Aguado (1985); Essery & Etchevers (2004)
OMEGAS1	Two stream parameter omega for snow [-]	0.8	0.64-0.96	0.802	--
OMEGAS2	Two stream parameter omega for snow [-]	0.4	0.32-0.48	0.402	--
RSURF_SNOW	Surface resistance for snow [s/m]	50.0	40.0-60.0	49.851	--
SNDECAYEXP	Exponent in snow decay albedo relationship [h^{-1}]	0.01	0.001-0.10	0.0338	Essery & Etchevers (2004)
SSI	Liquid water holding capacity for snowpack [m^3/m^3]	0.03	0.01-0.08	0.0398	Amorocho & Espildora (1966); Anderson (1973)
SWEMX	New snow mass to fully cover old snow [mm]	1.0	0.5-5.0	2.280	Xia et al. (2012)
T_LLIMIT	Lower temperature limit in rain-snow partitioning [C]	0.5	0.0-2.0	0.707	--
T_MLIMIT	Middle temperature limit in rain-snow partitioning [C]	2.0	0.5-3.0	1.724	--
T_ULIMIT	Upper temperature limit in rain-snow partitioning [C]	2.5	1.0-5.0	3.393	--

Z0SNO	Snow surface roughness length [m]	0.002	0.0001-0.01	0.00298	Marks & Dozier (1992); Reba et al. (2014)
SNOWF_SCALEF	Snowfall scale factor [-]	N/A	0.1-10.0	1.159	--

Table 2. Details of six evaluation points, including location, elevation, and percent tree canopy cover.

Evaluation Point	Latitude	Longitude	Elevation (m)	Tree Canopy Cover (%)*
Senator Beck	37.91° N	107.73° W	3721	14
Niwot	40.03° N	105.58° W	3185	79
Fool Creek	39.87° N	105.87° W	3400	89
Cameron Pass	40.52° N	105.89° W	3129	83
Rock Creek	38.98° N	107.03° W	3395	19
Skyway/Grand Mesa	39.05° N	108.06° W	3245	71

* Tree canopy cover calculated from Landsat 7 ETM+ data at 30 m spatial resolution (Hansen et al., 2013).

Table 3. Snow depth bias and RMSE for calibrated and uncalibrated Noah-MP simulations compared to UA and SNODAS for six SnowEx field site locations and the full western

Colorado domain. Bold indicates better performance, and for the overall domain comparisons, an asterisk (*) indicates a statistically significantly difference between the two model performances

Evaluation Point	UA				SNODAS			
	Noah-MP-Cal Snow Depth Bias (m)	Noah-MP-Def Snow Depth Bias (m)	Noah-MP-Cal Snow Depth RMSE (m)	Noah-MP-Def Snow Depth RMSE (m)	Noah-MP-Cal Snow Depth Bias (m)	Noah-MP-Def Snow Depth Bias (m)	Noah-MP-Cal Snow Depth RMSE (m)	Noah-MP-Def Snow Depth RMSE (m)
Senator Beck	0.247	**-0.183**	0.418	**0.315**	**0.163**	-0.267	**0.349**	0.428
Niwot	**-0.0663**	-0.245	**0.211**	0.387	**-0.0398**	-0.218	**0.279**	0.397
Fool Creek	-0.131	**-0.0769**	0.229	**0.207**	-0.327	**-0.119**	0.484	**0.227**
Cameron Pass	**-0.104**	-0.142	**0.211**	0.254	**-0.0267**	-0.0647	**0.170**	0.197
Rock Creek	**-0.180**	-0.251	**0.308**	0.400	**-0.187**	-0.259	**0.346**	0.428
Skyway/ Grand Mesa	**-0.0229**	-0.176	**0.173**	0.320	**-0.100**	-0.253	**0.246**	0.425
Overall Domain	**-0.00229***	-0.0362	**0.131***	0.146	**-0.00980***	-0.0437	0.186	**0.179***

Table 4. Comparison of seasonal bias and RMSE of snow depth for evaluation points from Noah-MP-Def and Noah-MP-Cal for the accumulation (December-February), peak snow (March-April), and ablation (May-July) seasons. Bold indicates better performance.

Point	Simulation	Accumulation Season Bias (cm)	Accumulation Season RMSE (cm)	Peak Season Bias (cm)	Peak Season RMSE (cm)	Ablation Season Bias (cm)	Ablation Season RMSE (cm)
Cameron Pass	Noah-MP-Def	-15.51	19.43	-30.35	33.86	-19.63	37.66
	Noah-MP-Cal	**-11.15**	**16.17**	**-24.77**	**29.12**	**-14.14**	**30.49**
Fool Creek	Noah-MP-Def	**7.05**	**15.10**	**15.60**	**29.76**	11.73	29.71
	Noah-MP-Cal	-17.89	23.37	-31.59	39.20	**-11.55**	**22.69**
Niwot Ridge	Noah-MP-Def	-34.03	40.62	-51.37	57.61	-26.02	45.72
	Noah-MP-Cal	**-8.08**	**21.59**	**-11.68**	**30.11**	**-10.71**	**26.11**
Rock Creek	Noah-MP-Def	-40.47	47.98	-64.26	68.27	**-15.64**	**31.85**
	Noah-MP-Cal	**-22.79**	**30.72**	**-46.41**	**51.51**	-17.59	32.96
Senator Beck	Noah-MP-Def	-28.24	**35.50**	-49.51	54.09	**-12.79**	**27.68**
	Noah-MP-Cal	**27.11**	36.84	**32.71**	**49.86**	41.22	61.32
Skyway (Grand Mesa)	Noah-MP-Def	-15.82	20.26	-48.90	55.45	-19.85	40.17
	Noah-MP-Cal	**4.42**	**12.22**	**-11.97**	**25.51**	**-6.20**	**24.41**

Table 5. Error metrics for snow depth and SWE comparing Noah-MP-Cal and Noah-MP-Def with snow pit observations from Grand Mesa and Senator Beck SnowEx study sites from the February 2017 field campaign. Bold indicates better performance between the two Noah-MP configurations.

		Snow Depth				SWE			
Study Site	Simulation	Mean Bias (cm)	MAE (cm)	Mean % diff (abs. value)	RMSE (cm)	Mean Bias (mm)	MAE (mm)	Mean % diff (abs. value)	RMSE (mm)
Grand Mesa	Noah-MP-Cal	**-12.1**	**28.1**	**20.0%**	**34.9**	**-23.0**	**106.4**	**23.6%**	**132.9**
	Noah-MP-Def	-48.2	48.8	32.2%	54.4	-160.6	162.9	32.6%	185.4
	SnowModel	-25.3	34.6	23.8%	41.1	-35.5	88.5	20.2%	112.2
Senator Beck	Noah-MP-Cal	92.6	92.6	84.4%	102.5	386.9	388.4	111.5 %	413.0
	Noah-MP-Def	**-13.1**	**43.1**	**34.0%**	**49.7**	**-32.1**	**142.5**	**37.6%**	**167.6**

Table 6. Nash-Sutcliffe Efficiency values for calibrated and uncalibrated Noah-MP simulations for six subbasins, as described by their USGS streamgage ID. Included in the number of Noah-MP grid cells within each subbasin. Bold indicates better performance. Asterisk indicates where monthly streamflow difference between Noah-MP-Cal and Noah-MP-Def is statistically significant at the 95% confidence level.

Basin	Basin ID	Noah-MP-Cal NSE	Noah-MP-Def NSE	Basin Area (km^2)
Colorado River at Glenwood Springs, CO	9072500	**0.56***	-0.065	11784
Taylor River below Taylor Park Reservoir, CO	9109000	**-1.03**	-1.96	656
Gunnison River below Blue Mesa Dam, CO	9124700	-3.62	**-0.71**	8933
Gunnison River below Crystal Reservoir, CO	9127800	-2.95	**-0.59**	10264
Crystal River above Avalanche Creek, CO	9081600 (CAMELS)	**0.72**	0.43	436
Taylor River at Taylor Park, CO	9107000 (CAMELS)	**-0.13**	-1.14	331

11. Figures

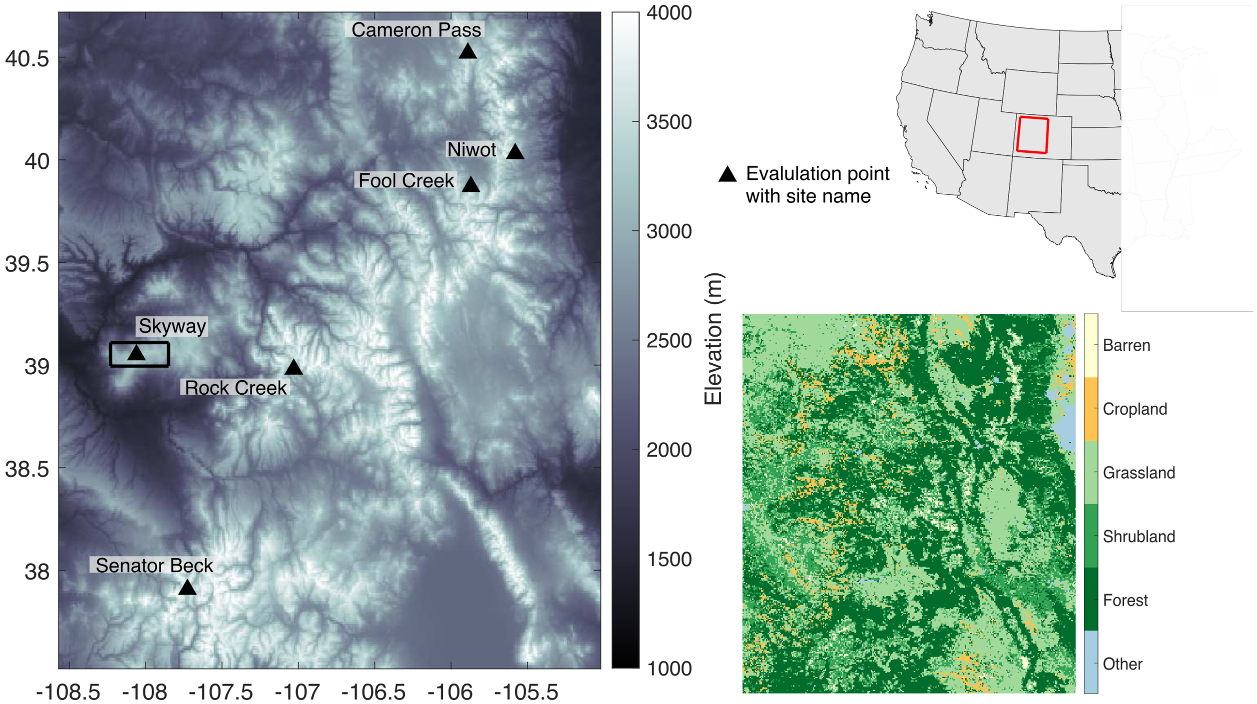

Figure 1. Elevations of western Colorado Noah-MP domain. Black box indicates the Grand Mesa intensive observation period field site from the NASA SnowEx 2017 field campaign. Triangles mark the six evaluation points and are labeled with the evaluation site name. The inset map shows the western Colorado domain with respect to the western United States. The bottom right plot shows the land classes for the model domain.

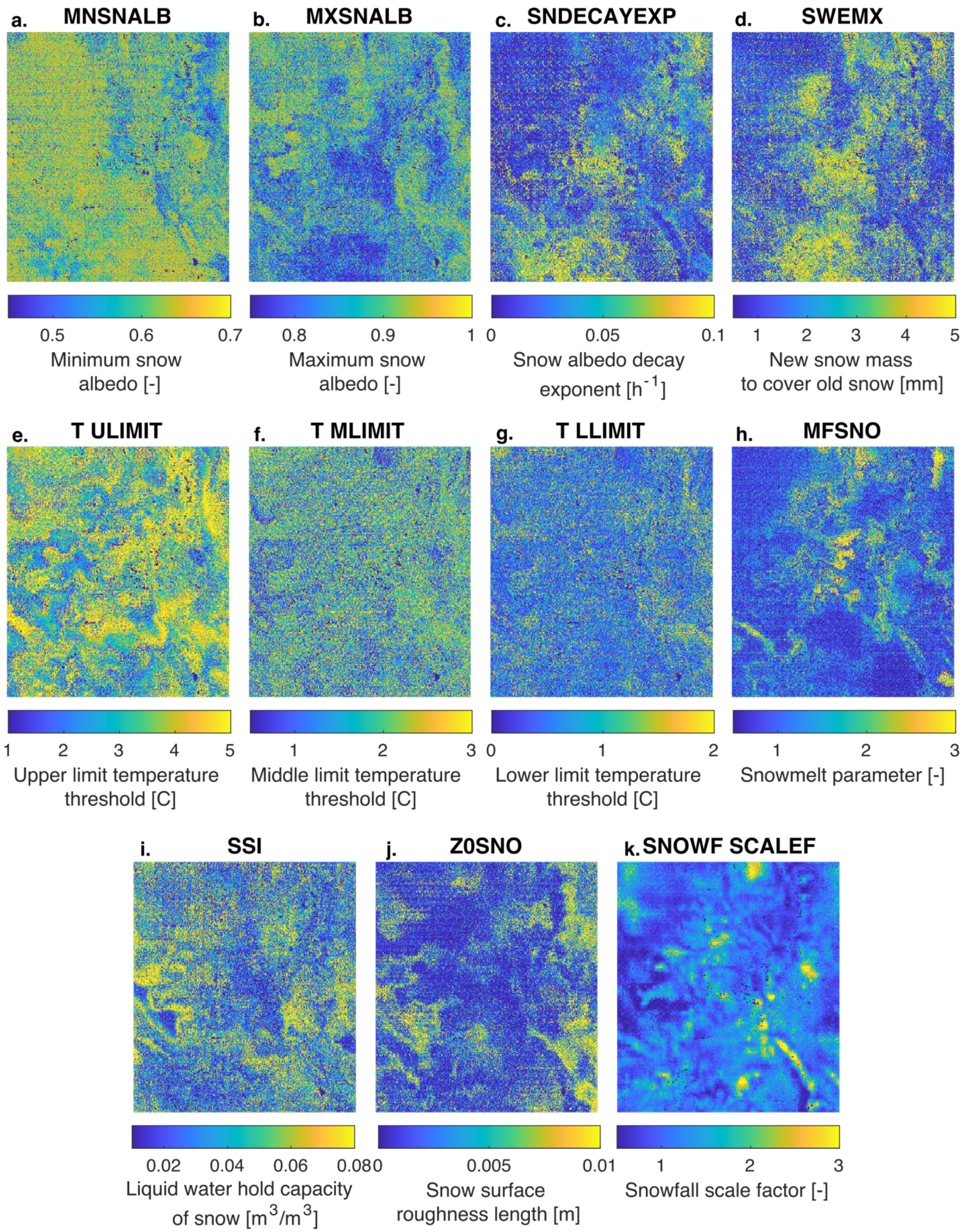

Figure 2. Calibrated parameters after the genetic algorithm procedure. Shown here are the 11 parameters that are most sensitive to calibration.

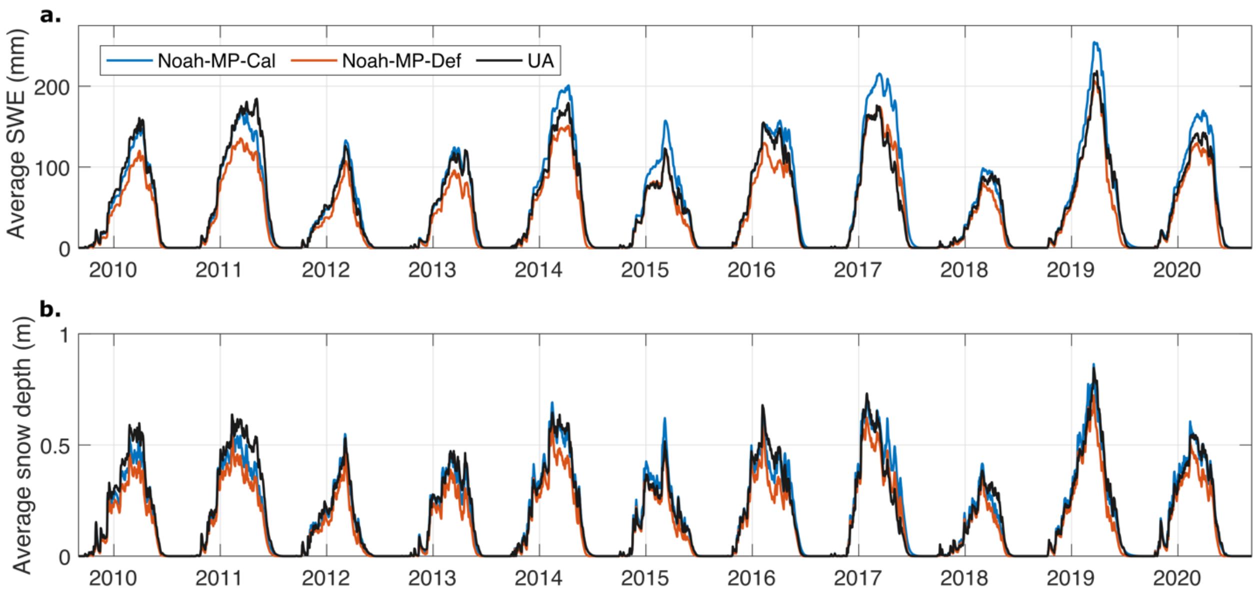

Figure 3. Time series of average SWE, in mm, and average snow depth, in m, over the full domain for calibrated (blue), uncalibrated (orange), and UA (black) estimates. Note that at the time of this analysis, UA ends in 2017 but the Noah-MP simulations continue through 2020.

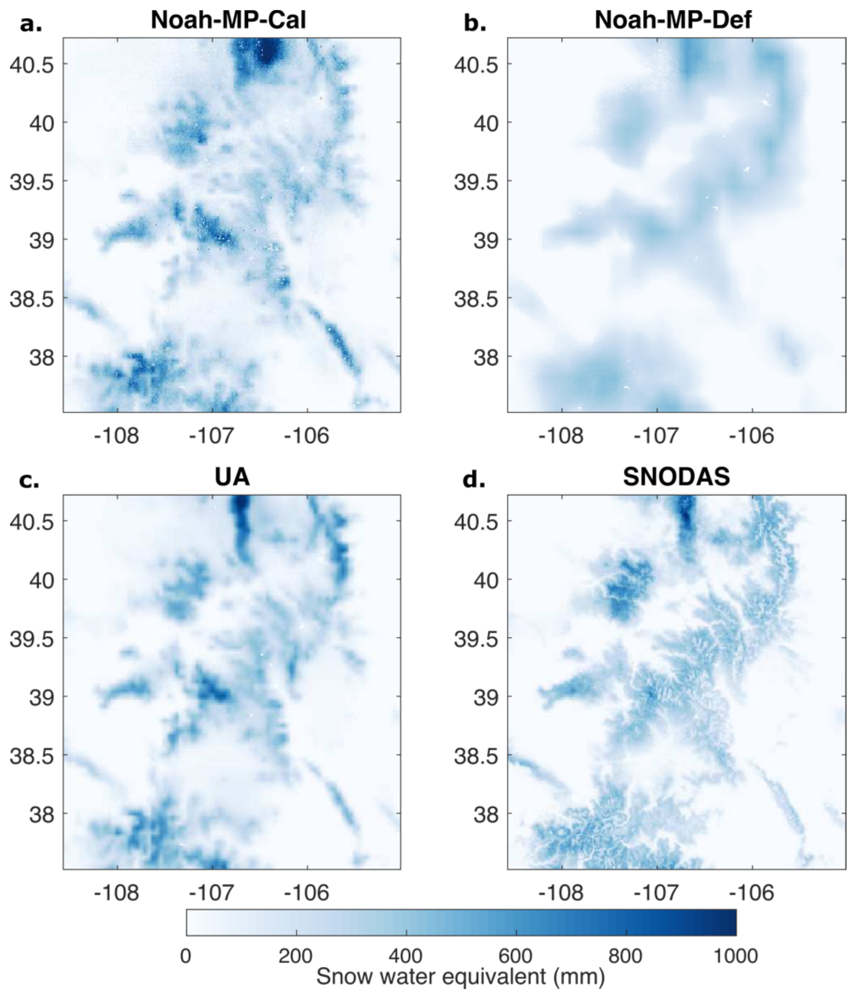

Figure 4. Average April 1 SWE, in mm, for the calibrated (a) and uncalibrated (b) simulations. (c) April 1 SWE difference, where blue indicates grid cells where the calibrated simulation has larger SWE and red indicates where the uncalibrated simulation has larger SWE.

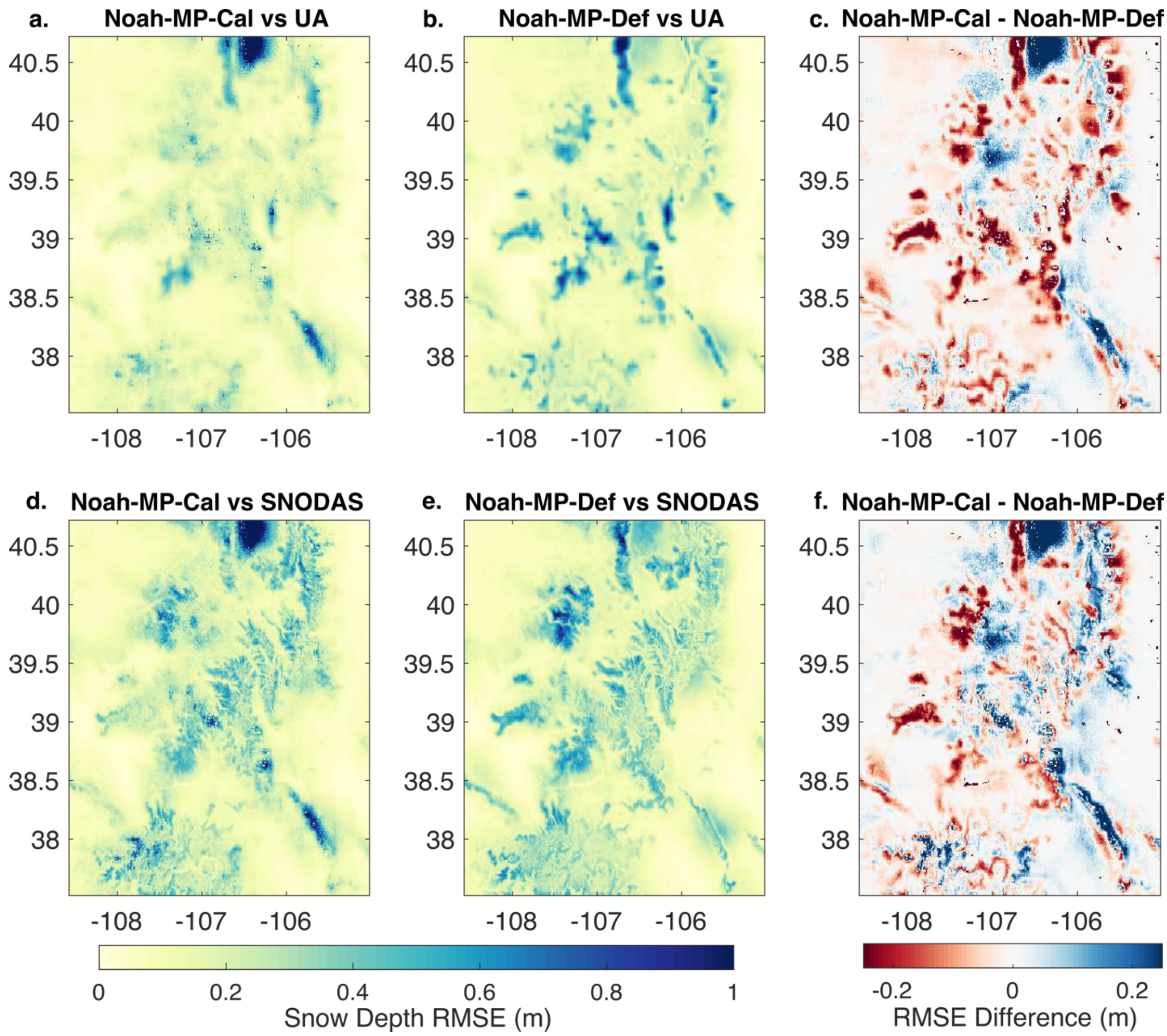

Figure 5. (top row) Snow depth RMSE for Noah-MP-Cal and Noah-MP-Def compared to UA for the full analysis period. The right column shows the difference in RMSE values between Noah-MP-Cal and Noah-MP-Def. (bottom row) Same as top row except compared against SNODAS.

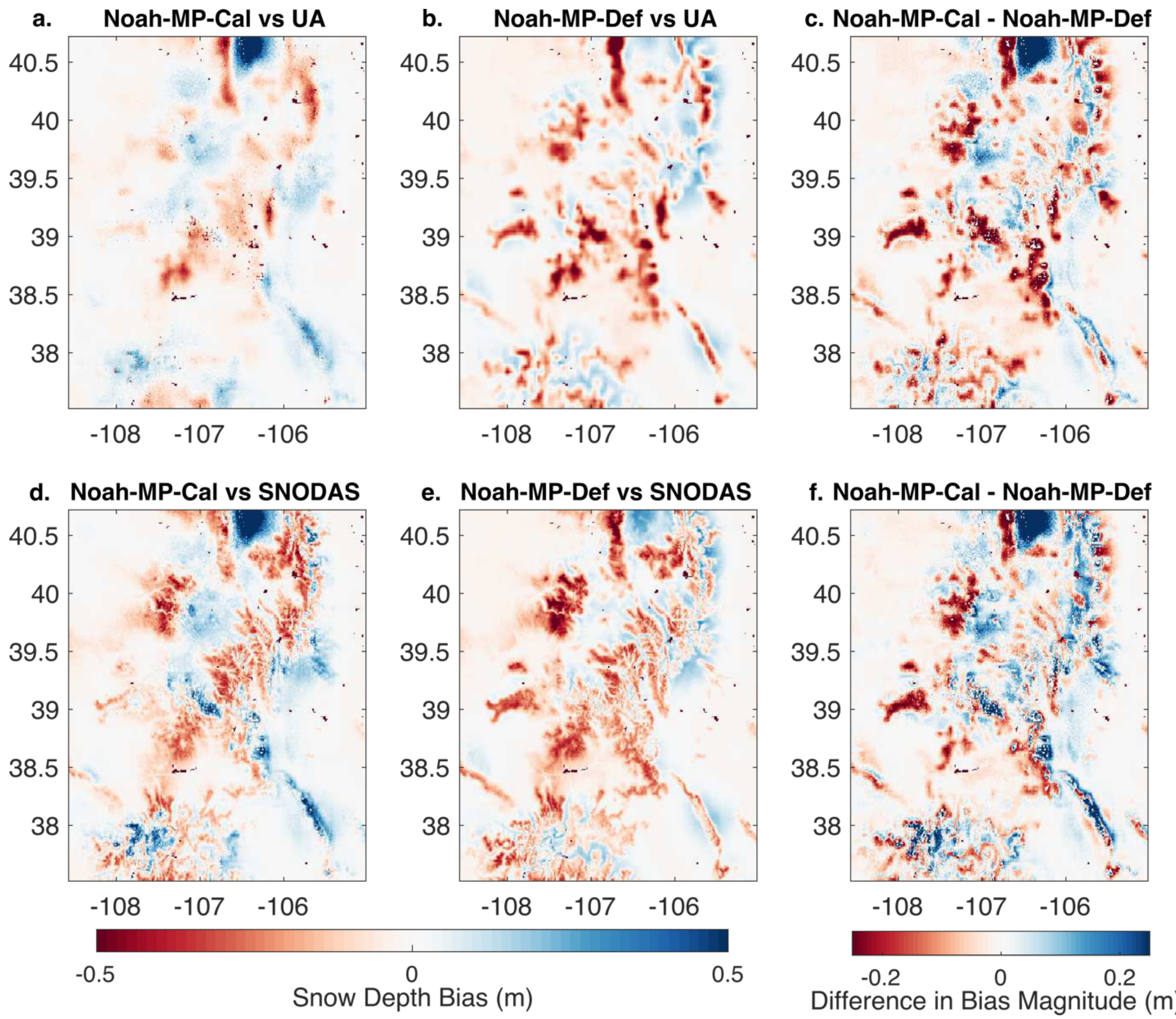

Figure 6. As for Figure 5 but for snow depth bias.

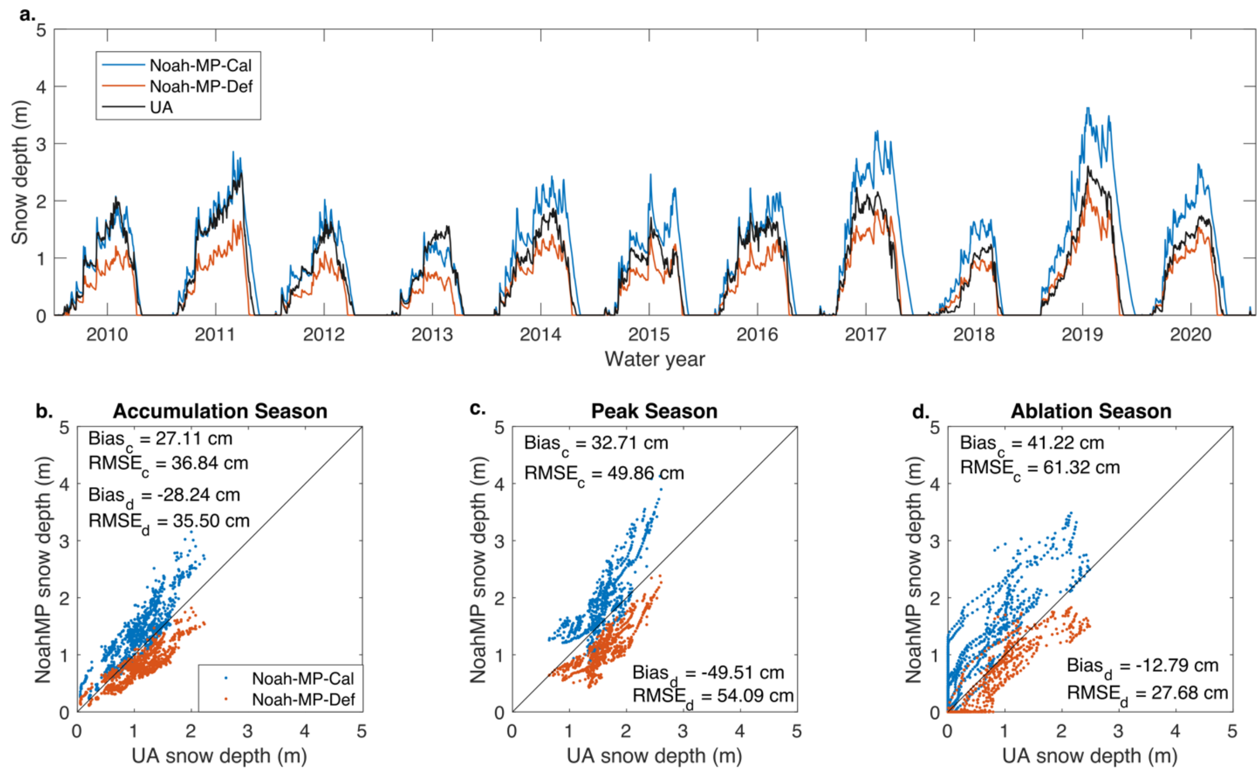

Figure 7. Evaluation of calibrated and uncalibrated Noah-MP over a single point in the Senator Beck basin. (a) Time series of daily snow depth over the grid cell that contains the Senator Beck study site. (bottom row) Scatter plot of Noah MP simulated snow depth verses UA snow depth for both calibrated (blue) and uncalibrated (orange) simulations, separated in accumulation (b; December-February), peak (c; March-April), and ablation (d; May-July) seasons.

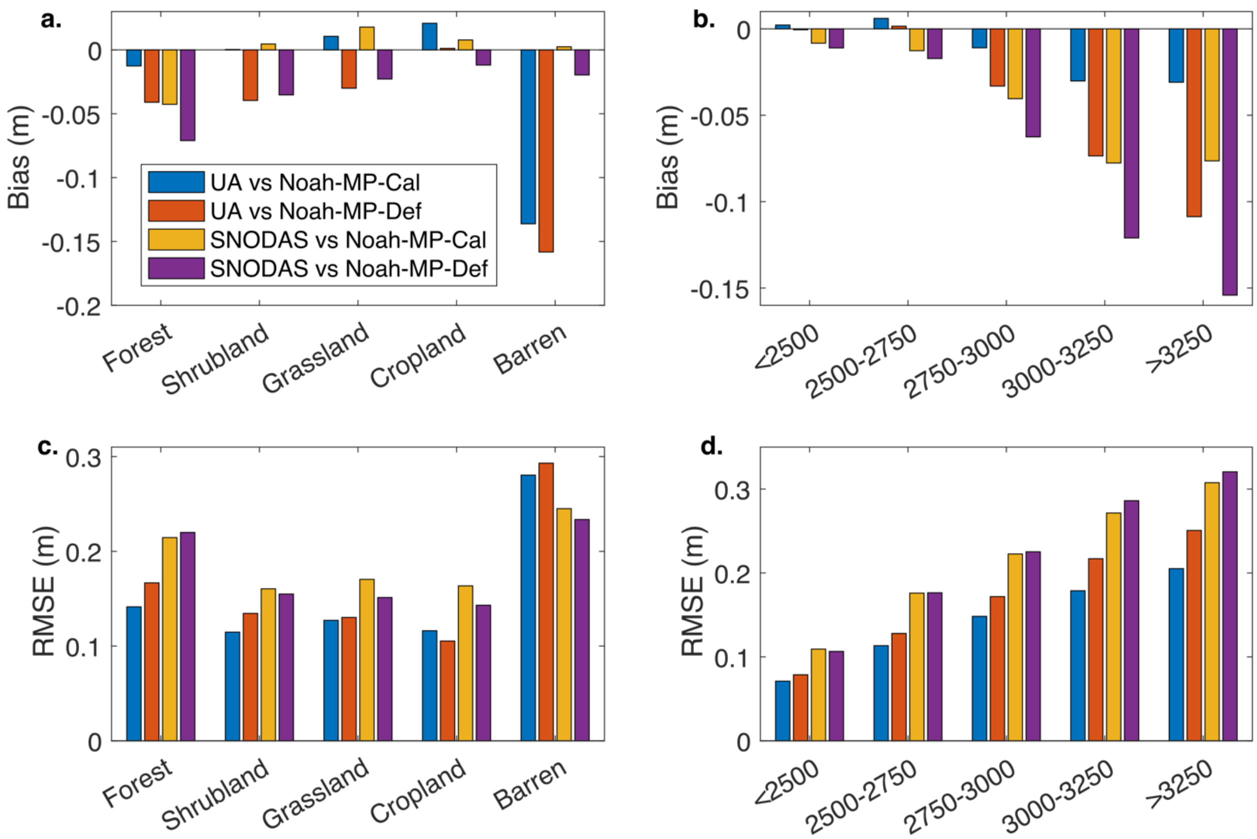

Figure 8. Snow depth bias (a) and RMSE (c) from Noah-MP-Cal and Noah-MP-Def compared to UA and SNODAS for five aggregated land cover categories. Snow depth bias (b) and RMSE (d) for the forest land cover category separated into elevation bands. Bias and RMSE values are temporal averages from the full analysis period.

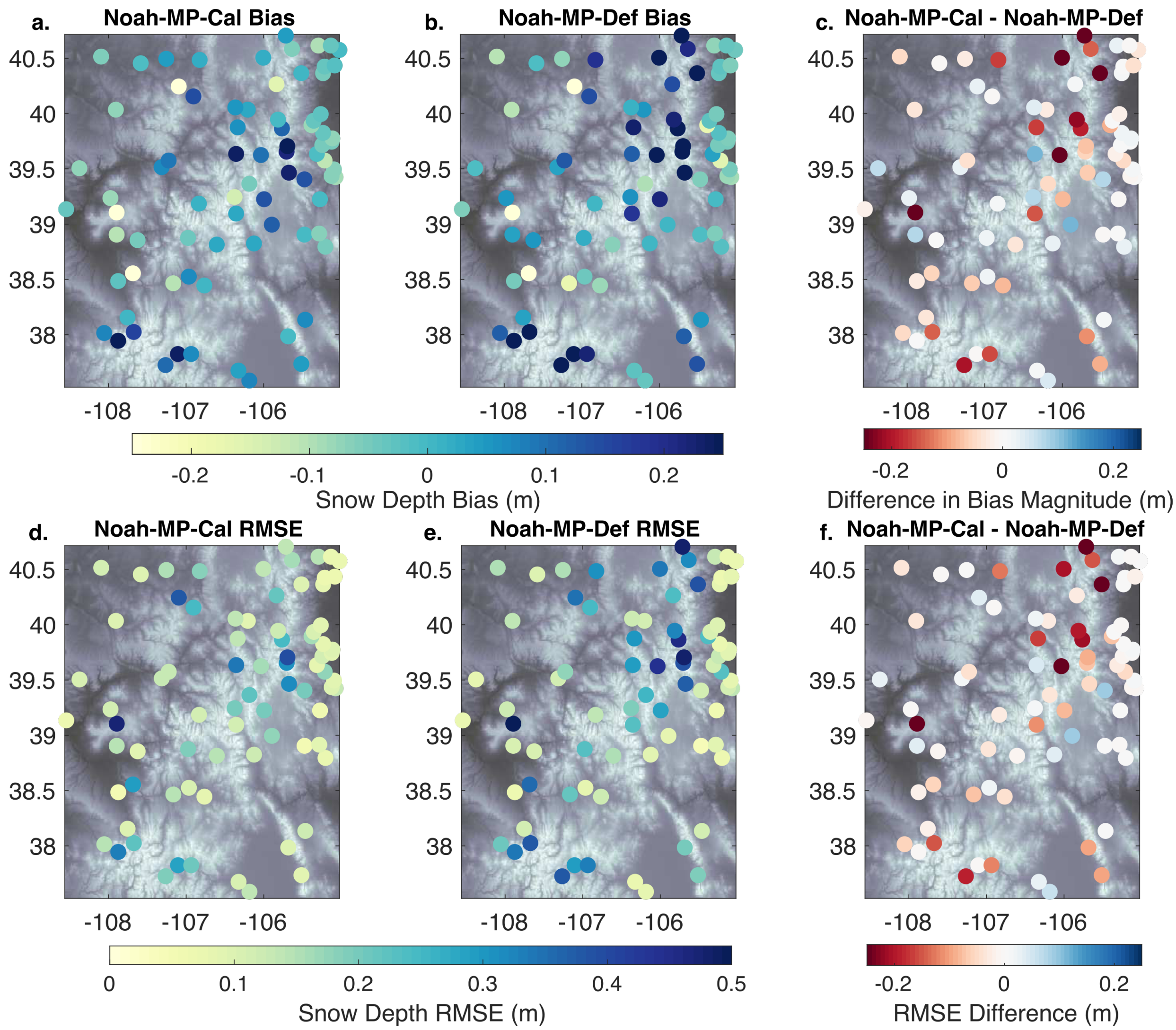

Figure 9. Snow depth bias (top row) and RMSE (bottom row) from Noah-MP-Cal and Noah-MP-Def compared to GHCN station observations. The right column shows the difference between Noah-MP-Cal and Noah-MP-Def for bias and RMSE.

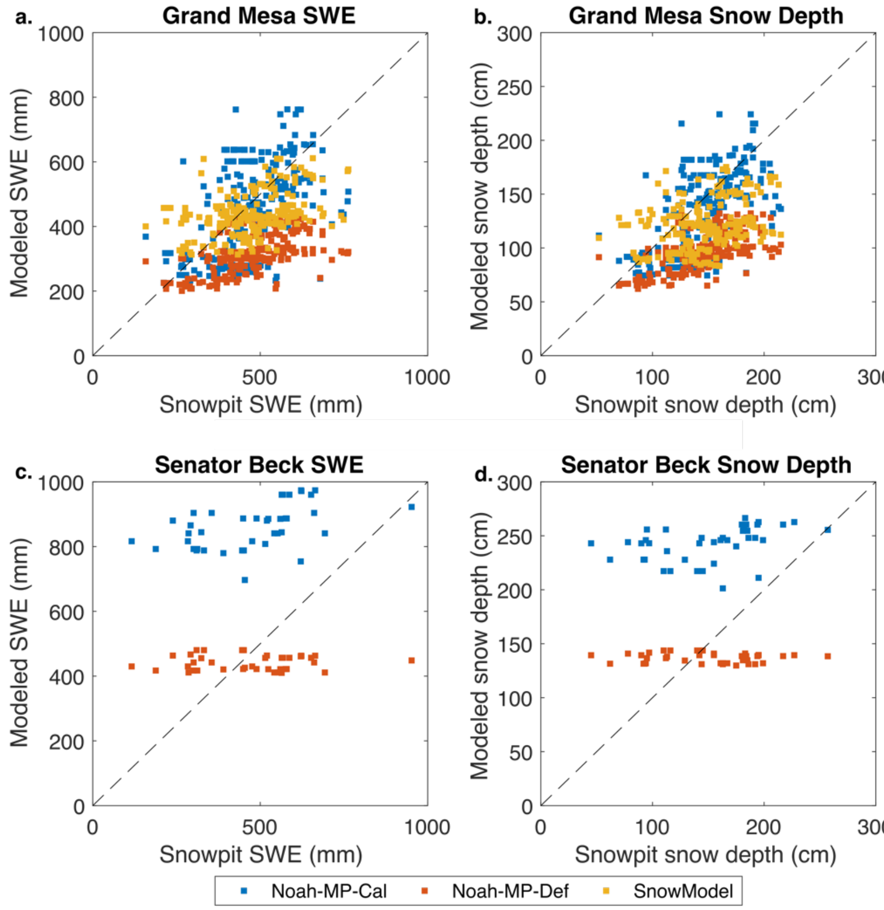

Figure 10. Comparison of SWE and snow depth between Noah-MP and SnowModel simulations and observations from snow pits during the SnowEx 2017 field campaign. In all plots, blue squares are calibrated Noah-MP, orange squares are default Noah-MP, and yellow squares are SnowModel (at native 30 m resolution). Plots (a) and (b) compare snow pit measurements for Grand Mesa and plots (c) and (d) compare for Senator Beck.

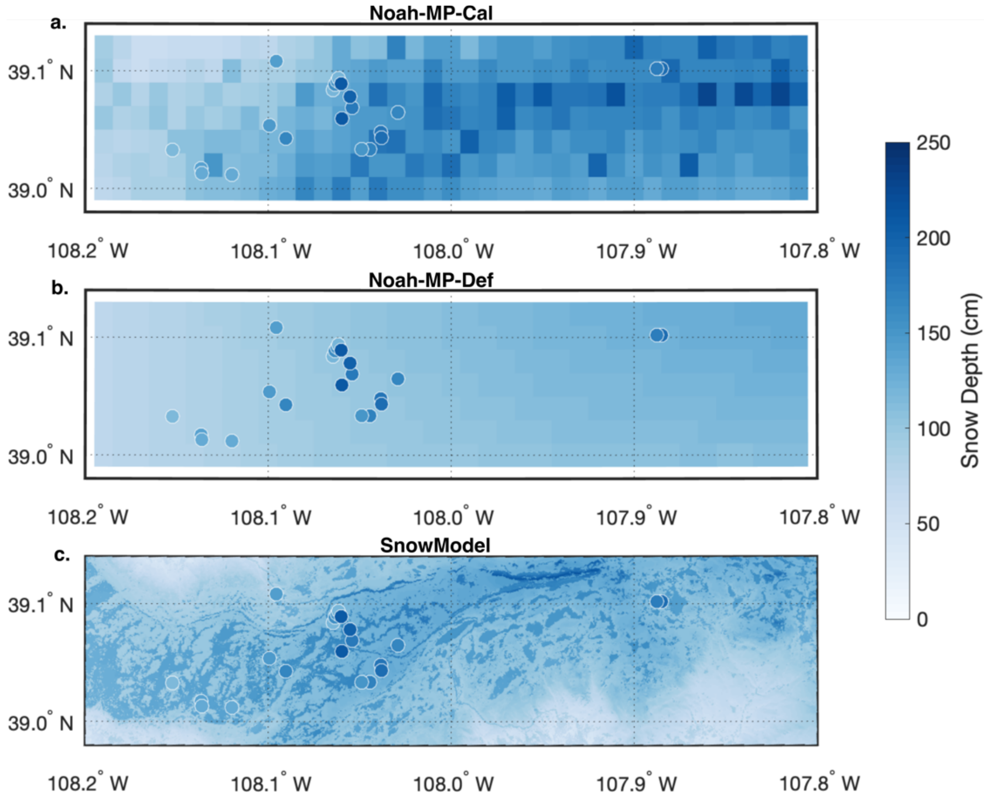

Figure 11. Comparison of Noah-MP-Cal (a), Noah-MP-Def (b), and SnowModel (c) snow depth estimates with snow pit observations for February 22, 2017, over the SnowEx Grand Mesa field campaign site. SnowModel is shown at 30 m spatial resolution. Snow pit depths and model depths are on the same color scale.

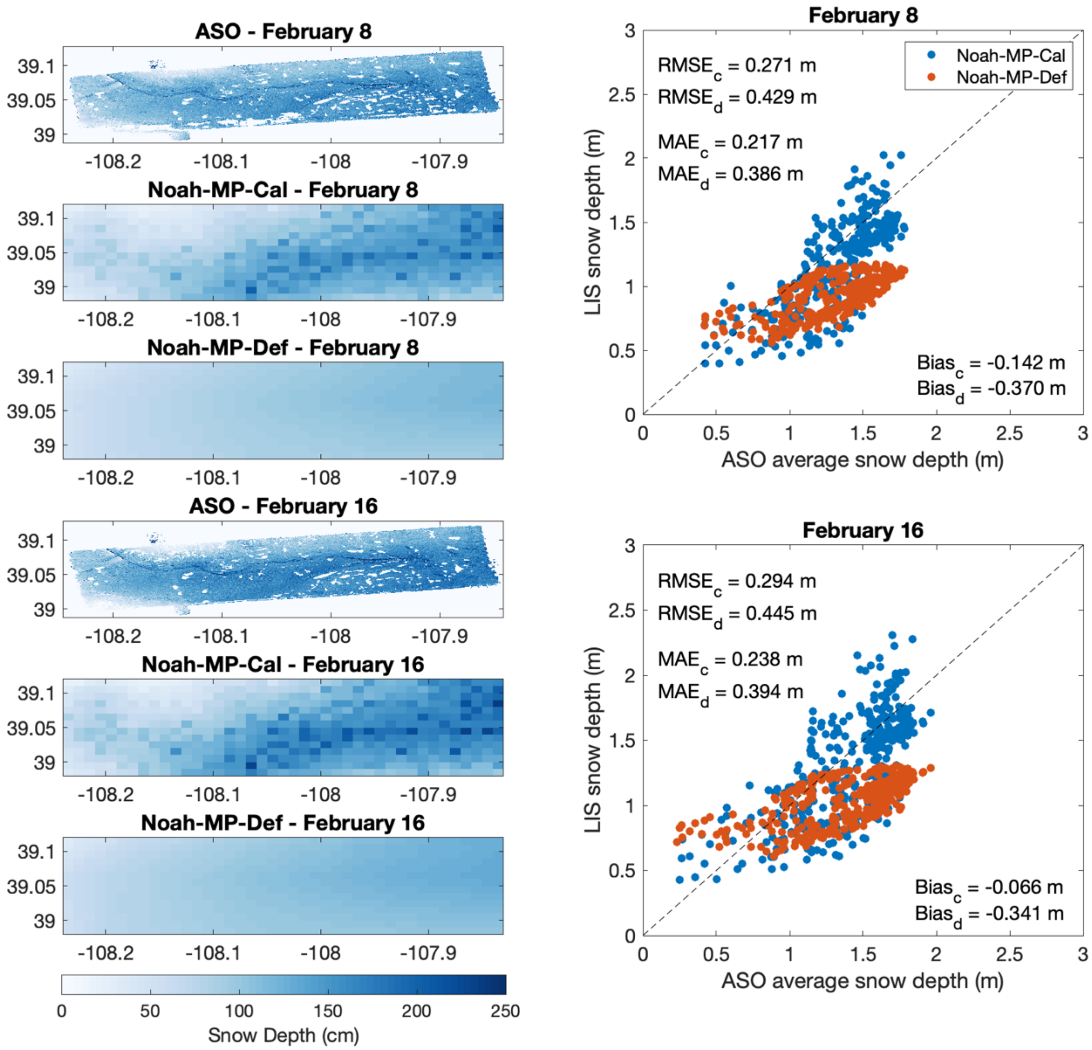

Figure 12. Comparison of snow depth from ASO flights with Noah-MP-Cal and Noah-MP-Def over Grand Mesa for February 8 and 16, 2017. Spatial maps are all at native resolution: 3 m for ASO and 0.01° for Noah-MP simulations. Scatter plots compare Noah-MP simulations to ASO observations, where ASO has been upscaled to 0.01° resolution.

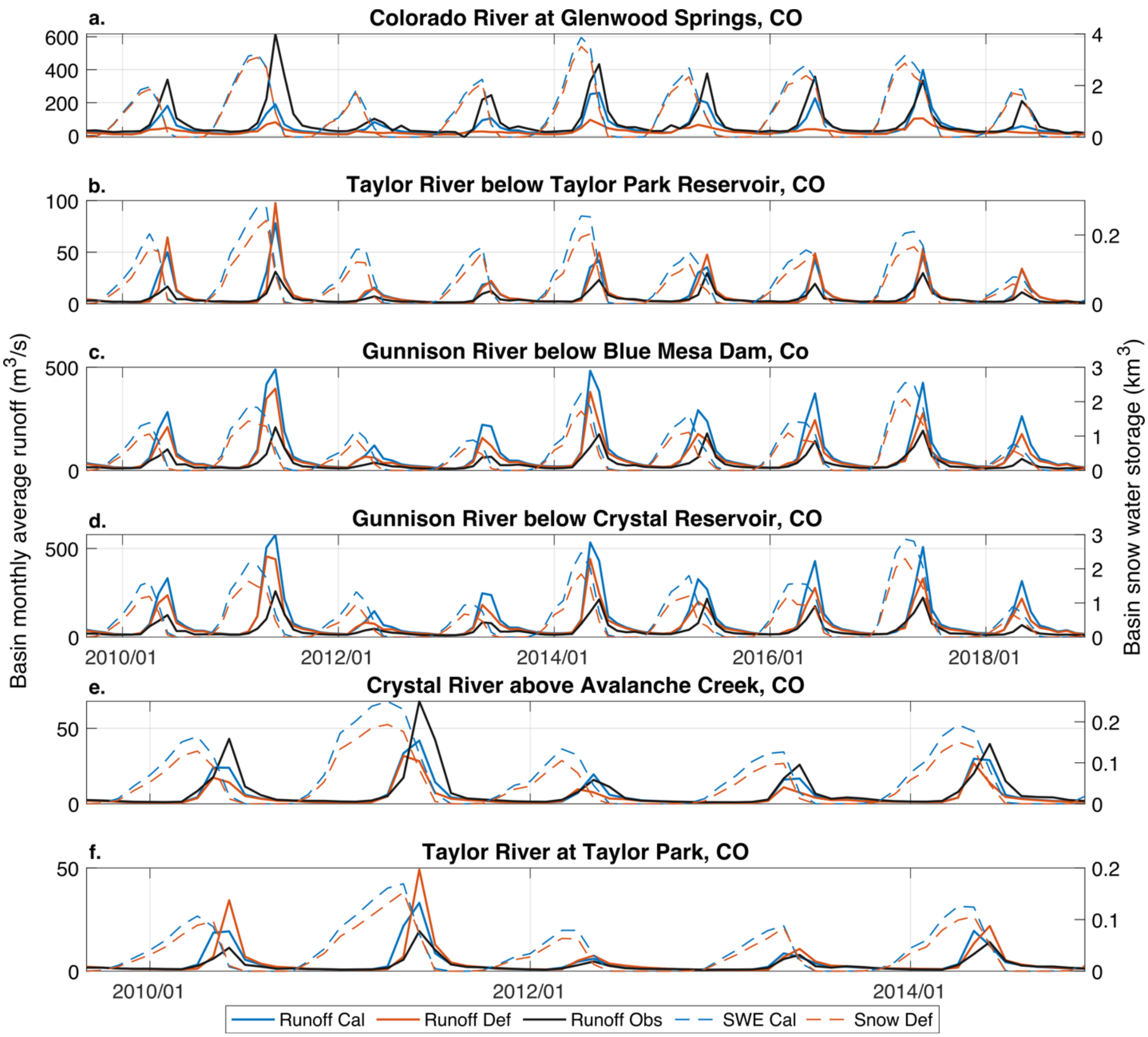

Figure 13. (a-d) Comparison of runoff, in m^3/month, between Noah-MP simulations and estimates of naturalized flow for four subbasins in the Upper Colorado River Basin. (e-f) Comparison of runoff, in m^3/month, between Noah-MP simulations and observed streamflow from USGS streamgages for small unmanaged subbasins, selected from the CAMELS database. Streamgage locations are shown on Figure S4. Dashed lines in all plots show basin snow water storage, in km^3.